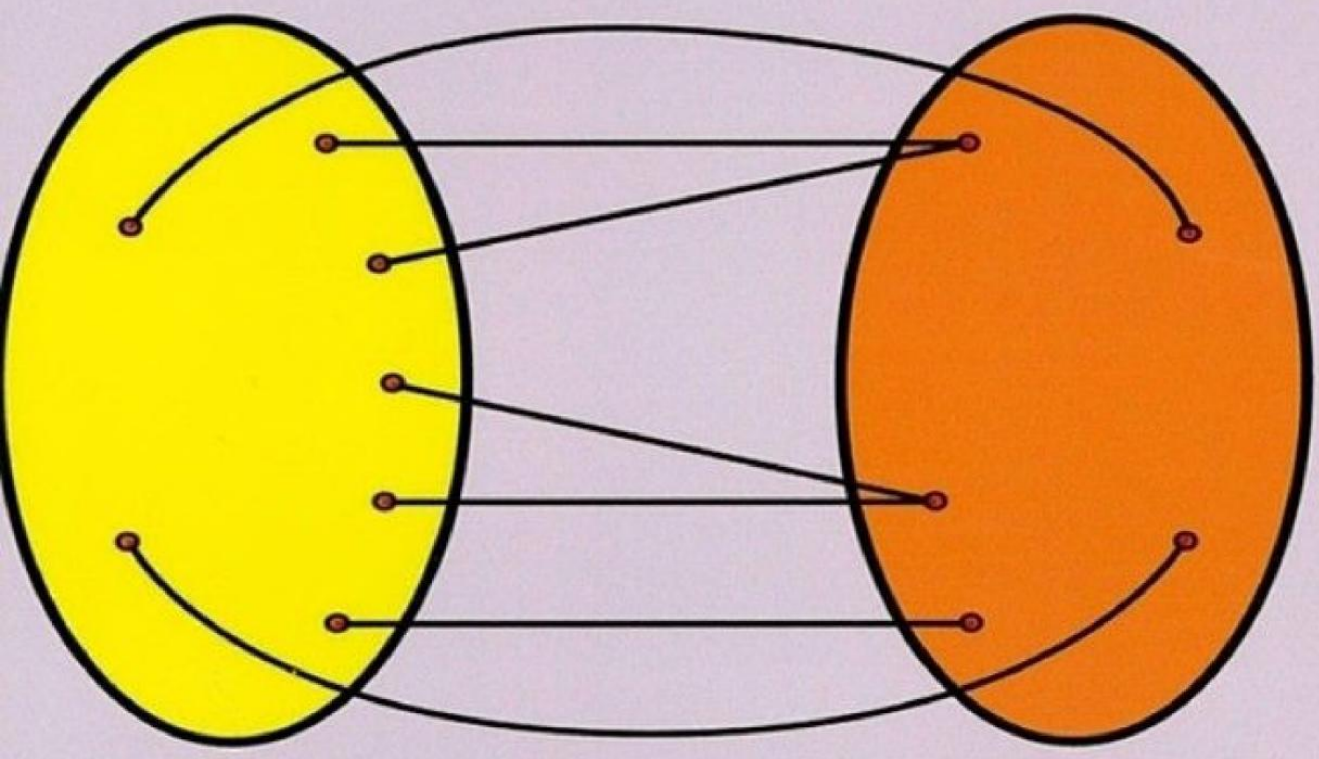
AN INTRODUCTION TO
Information Theory



Fazlollah M. Reza

Bibliographical Note

This Dover edition, first published in 1994, is an unabridged and unaltered republication of the work first published by the McGraw–Hill Book Company, Inc., New York, in 1961 in the *McGraw–Hill Electrical and Electronic Engineering Series.*

Library of Congress Cataloging–in–Publication Data

Reza, Fazlollah M.

An introduction to information theory / Fazlollah M. Reza.

p. cm.

Originally published: New York: McGraw–Hill, 1961.

Includes bibliographical references and index.

9780486158440

1. Information theory. 2. Probabilities. I. Title.

Q360.R43 1994
003'.54—dc20

94–27222

CIP

Manufactured in the United States by Courier Corporation
68210205
www.doverpublications.com

Table of Contents

CHAPTER 1

INTRODUCTION

Information theory is a new branch of probability theory with extensive potential applications to communication systems. Like several other branches of mathematics, information theory has a physical origin. It was initiated by communication scientists who were studying the statistical structure of electrical communication equipment.

Our subject is about a decade old. It was principally originated by Claude Shannon through two outstanding contributions to the mathematical theory of communications in 1948 and 1949. These were followed by a flood of research papers speculating upon the possible applications of the newly born theory to a broad spectrum of research areas, such as pure mathematics, radio, television, radar, psychology, semantics, economics, and biology. The immediate application of this new discipline to the fringe areas was rather premature. In fact, research in the past 5 or 6 years has indicated the necessity for deeper investigations into the foundations of the discipline itself.

Despite this hasty generalization which produced several hundred research papers (with frequently unwarranted conclusions), one thing became evident. The new scientific discovery has stimulated the interest of thousands of scientists and engineers around the world.

Our first task is to present a bird's–eye view of the subject and to specify its place in the engineering curriculum. In this chapter a heuristic exposition of the topic is given. No effort is made to define the technical vocabulary. Such an undertaking requires a detailed logical presentation and is out of place in this informal introduction. However, the reader will find such material presented in a pedagogically prepared sequence beginning with Chap. 2. This introductory chapter discusses generalities, leaving a more detailed and precise treatment to

subsequent chapters.[1] The specialist interested in more concrete statements may wish to forgo this introduction and begin with the body of the book.[1]

1-1. Communication Processes.

Communication processes are concerned with the flow of some sort of information–carrying commodity in some *network.* The commodity need not be tangible; for example, the process by which one mind affects another mind is a communication procedure. This may be the sending of a message by telegraph, visual communication from artist to viewer, or any other means by which *information* is conveyed from a transmitter to a receiver. The subject matter deals with the gross aspects of communication models rather than with their minute structure. That is, we concentrate on the over–all performance of such systems without being restrained to any particular equipment or organ. Common to all communication processes is the flow of some commodity in some network. While the nature of the commodity can be as varied as electricity, words, pictures, music, and art, one could suggest at least three essential parts of a communication system (Fig. 1-1):

1. Transmitter or source

2. Receiver or sink

3. Channel or transmission network which conveys the communiqué from the transmitter to the receiver

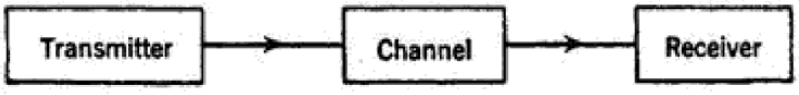


FIG. 1-1. The model of a communication system.

This is the simplest communication system that one can visualize. Practical cases generally consist of a number of sources and receivers and a complex network. A familiar analogous example is an electric power system using several interconnected power plants to supply several towns.

In such problems one is concerned with a study of the distribution of the commodity in the network, defining some sort of efficiency of transmission and hence devising schemes leading to the most efficient transmission.

When the communiqué is tangible or readily measurable, the problems encountered in the study of the communication system are of the types somewhat familiar to engineers and operational analysts (for instance, the study of an electric circuit or the production schedule of a manufacturing plant). When the communiqué is "intelligence" or "information," this general familiarity cannot be assumed. How does one define a measure for the amount of *information?* And having defined a suitable measure, how does one apply it to the betterment of the communication of information?

To mention an analog, consider the case of an electric power network transmitting electric energy from a source to a receiver (Fig. 1-2). At the source the electric energy is produced with voltage V_s. The receiver requires the electric energy at some prescribed voltage V_r. One of the problems involved is the heat loss in the channel (transmission line). In other words, the impedance of the wires acts as a parasitic receiver. One of the many tasks of the designer is to minimize the loss in the transmission lines. This can be accomplished partly by improving the quality of the transmission lines. A parallel method of transmission improvement is to increase the voltage at the input terminals of the line. As is well known, this improves the efficiency of transmission by reducing energy losses in the line. A step–up voltage transformer installed at the input terminals of the line is appropriate. At the output terminals another transformer (step–down) can provide the specified voltage to the receiver.

Without being concerned about mathematical discipline in this introductory chapter, let us ask if similar procedures could be applied to the transmission of *information*. If the channel of transmission of information is a lossy one, can one still improve the *efficiency* of the transmission by procedures similar to those in the above case? This of course depends, in the first place, on whether a measure for the efficiency of transmission of information can be defined.

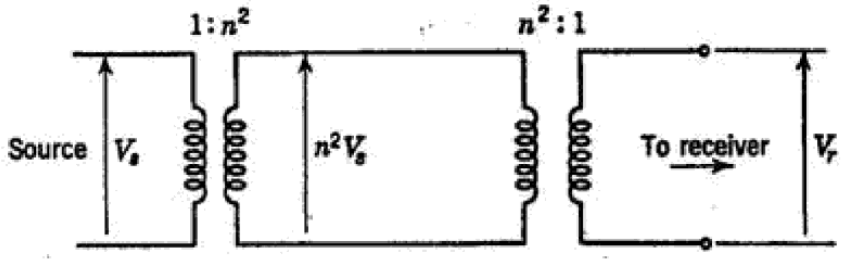


FIG. 1-2. An example of a communication system.

1-2. A Model for a Communication System.

The communication systems considered here are of a statistical nature. That is, the performance of the system can never be described in a deterministic sense; rather, it is always given in statistical terms. A source is a device that selects and transmits sequences of symbols from a given alphabet. Each selection is made at random, although this selection may be based on some statistical rule. The channel transmits the incoming symbols to the receiver. The performance of the channel is also based on laws of chance. If the source transmits a symbol, say A, with a probability of $P\{A\}$ and the channel lets through the letter A with a probability denoted by $P\{A|A\}$, then the probability of transmitting A and receiving A is

$$P\{A\} \cdot P\{A|A\}$$

The communication channel is generally lossy; i.e., a part of the transmitted commodity does not reach its destination or it reaches the destination in a distorted form. There are often unwanted sources in a communication channel, such as *noise* in radio and television or passage of a vehicle in the opposite direction in a one-way street. These sources of disturbance are generally referred to as *noise sources* or simply *noise.* An important task of the designer is the minimization of the loss and the optimum recovery of the original commodity when it is corrupted by the effect of noise.

In the deterministic electrical model of Fig. 1-2, it was pointed out that one

device which may be used to improve the efficiency of the system is called a *transformer.* In the vocabulary of information theory a device that is used to improve the efficiency of the channel may be called an *encoder.* An encoded message is less susceptible to channel noise. At the receiver's terminal a *decoder* is employed to transform the encoded messages into the original form which is acceptable to the receiver. It could be said that, in a certain sense, for more "efficient" communication, the encoder performs a one-to-one mathematical mapping or an operation F on the input commodity I, $F(I)$, while the decoder performs the inverse of that operation, F^{-1}.

Encoder:	F	I	$F(I)$
Decoder:	F^{-1}	$F(I)$	I

(1-1)

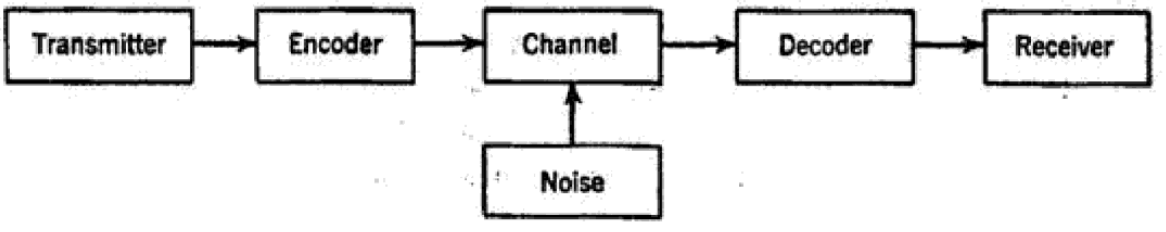


FIG. 1-3. General structure of a communication system used in information theory.

This perfect procedure is, of course, hypothetical; one has to face the ultimate effect of noise which in physical systems will prevent perfect communication. This is clearly seen in the case of the transmission of electrical energy where the transformer decreases the heat loss but an efficiency of 100 per cent cannot be expected. The step-up transformer acts as a sort of encoder and the step-down transformer as a decoding apparatus.

Thus, in any practical situation, we have to add at least three more basic parts to our mathematical model: source of noise, encoder, and decoder (Fig. 1-3). The model of Fig. 1-3 is of a general nature; it may be applied to a variety of circumstances.

A novel application of such a model was made by Wiener and Shannon in their discussions of the statistical nature of the communication of messages. It was pointed out that a radio, television, teletype, or speech transmitter selects sequences of messages from a known transmitter vocabulary *at random* but with specified probabilities. Therefore, in such communication models, the source, channel, encoder, decoder, noise source, and receiver must be *statistically* defined. This point of view in itself constitutes a significant contribution to the communication sciences. In light of this view, one comes to realize that a basic study of communication systems requires some knowledge of probability theory. Communication theories cannot be adequately studied without having a good background of probability. Conversely, readers acquainted with the fundamentals of probability theory can proceed most efficiently with research in the field of communication.

In the macroscopic study of communication systems, some of the basic questions facing us are these:

1. How does one measure information and define a suitable unit for such measurements?

2. Having defined such a unit, how does one define an information *source*, or how does one measure the rate at which an information source supplies information?

3. What is the concept of channel? How does one define the rate at which a *channel* transmits information?

4. Given a source and a channel, how does one study the joint rate of transmission of information and how does one go about improving that rate? How far can the rate be improved?

5. To what extent does the presence of noise limit the rate of transmission of information without limiting the communication reliability?

To present systematic answers to these questions is our principal task. This is undertaken in the following chapters. However, for the benefit of those who wish to acquire a heuristic introduction to the subject, we include a brief discussion of it here.

1-3. A Quantitative Measure of Information.

In our study we deal with ideal mathematical models of communication. We confine ourselves to models that are statistically defined. That is, the most significant feature of our model is its unpredictability. The source, for instance, transmits at random any one of a set of prespecified messages. We have no specific knowledge as to which message will be transmitted next. But we know the probability of transmitting each message directly, or something to that effect. If the behavior of the model were predictable (deterministic), then recourse to measuring an amount of information would hardly be necessary.

When the model is statistically defined, while we have no concrete assurance of its detailed performance, we are able to describe, in a sense, its "over-all" or "average" performance in the light of its statistical description. In short, our search for an amount of information is virtually a search for a statistical parameter associated with a probability scheme. The parameter should indicate a relative measure of uncertainty relevant to the occurrence of each particular message in the message ensemble.

We shall illustrate how one goes about defining the amount of information by a well-known rudimentary example. Suppose that you are faced with the selection of equipment from a catalog which indicates n distinct models:

$$[x_1, x_2, \ldots, x_n]$$

The desired amount of information $I(x_k)$ associated with the selection of a particular model x_k must be a function of the probability of choosing x_k:

$$I(x_k) = f(P\{x_k\}) \tag{1-2}$$

If, for simplicity, we assume that each one of these models is selected with an

equal probability, then the desired amount of information is only a function of n.

$$I_1(x_k) = f\left(\frac{1}{n}\right)$$

(1-2a)

Next assume that each piece of equipment listed in the catalog can be ordered in one of m distinct colors. If for simplicity we assume that the selection of colors is also equiprobable, then the amount of information associated with the selection of a color c_j among all equiprobable colors $[c_1, c_2, \ldots, c_m]$ is

$$I_2(c_j) = f(P\{c_j\}) = f\left(\frac{1}{m}\right)$$

(1-2b)

where the function $f(x)$ must be the same unknown function used in Eq. (1-2*a*).

Finally, assume that the selection is done in two ways:

1. Select the equipment and then select the color, the two selections being independent of each other.

2. Select the equipment and its color at the same time as one selection from mn possible equiprobable choices.

The search for the function $f(x)$ is based on the intuitive choice which requires the equality of the amount of information associated with the selection of the model x_k with color c_j in both schemes (1-2*c*) and (1-2*d*).

$$I(x_k \text{ and } c_j) = I_1(x_k) + I_2(c_j) = f\left(\frac{1}{n}\right) + f\left(\frac{1}{m}\right)$$

(1-2c)

$$I(x_k \text{ and } c_j) = f\left(\frac{1}{mn}\right)$$

(1-2d)

Thus

$$f\left(\frac{1}{n}\right) + f\left(\frac{1}{m}\right) = f\left(\frac{1}{mn}\right)$$

(1-3)

This functional equation has several solutions, the most important of which, for our purpose, is

$$f(x) = -\log x$$

(1-4)[2]

To give a numerical example, let $n = 18$ and $m = 8$.

$$\begin{aligned} I_1(x_k) &= \log 18 \\ I_2(c_j) &= \log 8 \\ I(x_k \text{ and } c_j) &= I_1(x_k) + I_2(c_j) \\ I(x_k \text{ and } c_j) &= \log 18 + \log 8 = \log 144 \end{aligned}$$

Thus, when a statistical experiment has n equiprobable outcomes, the average amount of information associated with an outcome is log n. The logarithmic information measure has the desirable property of additivity for independent

statistical experiments. These ideas will be elaborated upon in Chap. 3.

1-4. A Binary Unit of Information.

The simplest case to consider is a selection between two equiprobable events E_1 and E_2. E_1 and E_2 may be, say, head or tail in a throwing of an "honest" coin. Following Eq. (1-4), the amount of information associated with the selection of one out of two equiprobable events is

$$-\log \tfrac{1}{2} = \log 2$$

An arbitrary but convenient choice of the base of the logarithm is 2. In that case, $-\log_2 \frac{1}{2} = 1$ provides a unit of information. This unit is commonly known as a *bit.*[3]

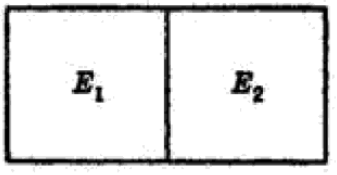


FIG. 1-4. A probability space with two equiprobable events.

Next consider the selection of one out of 2^2, 2^3, 2^4, . . . , 2^N equally likely choices. By successively partitioning a selection into two equally likely selections, we come to the conclusion that the amounts of information associated with the previous selection schemes are, respectively, 2, 3, 4, . . . , N bits.

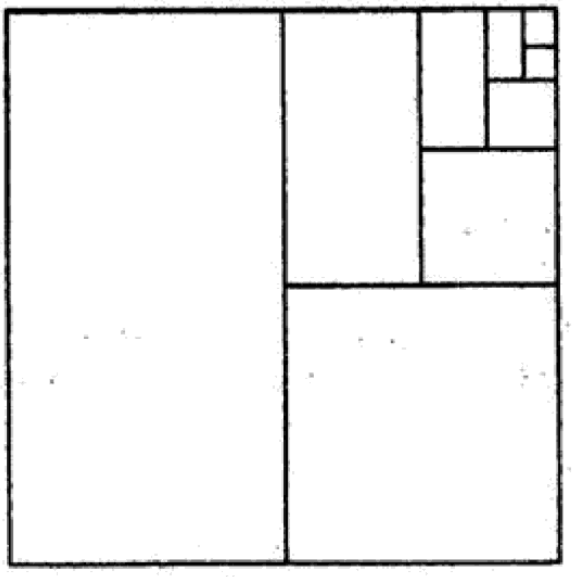


FIG. 1-5. Successive partitioning of the probability space.

In a slightly more general case, consider a source with a finite number of messages and their corresponding transmission probabilities.

$$\begin{matrix}[x_1, x_2, \ldots, x_n] \\ [P\{x_1\}, P\{x_2\}, \ldots, P\{x_n\}]\end{matrix}$$

The source selects at random each one of these messages. Successive selections are assumed to be *statistically independent.* The probability associated with the selection of message x_k is $P\{x_k\}$. The amount of information associated with the transmission of message x_k is defined as

$$I_k = -\log P\{x_k\}$$

I_k is also called the amount of self–information of the message x_k. The average information per message for the source is

$$I = \text{statistical average of } I_k = -\sum_{k=1}^{n} P\{x_k\} \log P\{x_k\}$$

(1-5)

For instance, the amount of information associated with a source of the above type, transmitting two symbols 0 and 1 with equal probability, is

$$I = -(\tfrac{1}{2} \log \tfrac{1}{2} + \tfrac{1}{2} \log \tfrac{1}{2}) = 1 \text{ bit}$$

If the two symbols were transmitted with probabilities and 1 - , then the average amount of information per symbol becomes

$$I = -\alpha \log \alpha - (1 - \alpha) \log (1 - \alpha)$$

(1-6)

The average information per message I is also referred to as the *entropy* (or the communication entropy) of the source and is usually denoted by the letter H. For instance, the entropy of a simple source of the above type is

$$H(p_1,p_2, \ldots ,p_n) = -(p_1 \log p_1 + p_2 \log p_2 + \ldots + p_n \log p_n)$$

where $(p_1,p_2, \ldots ,p_n)$ refers to a *discrete complete* probability scheme. Figure 1-6 shows the entropy of a simple binary source for different message probabilities.

Next, consider a second similar source having m symbols, and designate the amount of information per symbol of the two sources by $H(n)$ and $H(m)$, respectively. If the two sources transmit their symbols independently, their joint output might be considered as a source having mn distinct pairs of symbols. It

can be shown that for two such independent sources the average information per joint symbol is

$$H(mn) = H(m) + H(n)$$

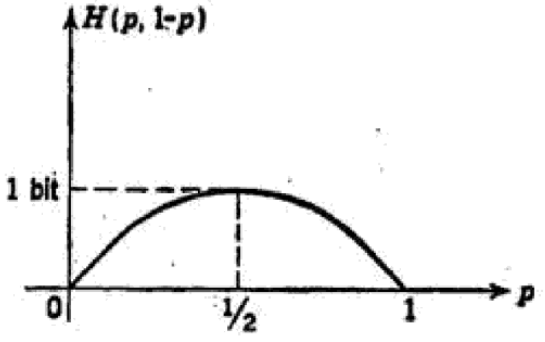


FIG.1-6. The entropy of an independent discrete memoryless binary source

The formal derivation of this relation is given in Chap. 3.

1-5. Sketch of the Plan.

From a mathematical point of view, the heuristic exposition of the previous two sections is somewhat incomplete.

We still need to formalize our understanding of the basic concepts involved and to develop techniques for studying more complex physical models. It was suggested that, given an *independent source S* which transmits messages x_k from a finite set

$$\begin{matrix}[x_1, x_2, \ldots, x_n] \\ [P\{x_1\}, P\{x_2\}, \ldots, P\{x_n\}]\end{matrix}$$

there is an average amount of information $I(x)$ associated with the independent source S.

$I(x)$ = expected value or average of $I(x_k)$ for all messages

Our next step is to generalize this to the case of random variables with two or more not necessarily statistically independent dimensions, for instance, to define the amount of information per symbol of a scheme having pairs of statistically related symbols (x_k,y_k). This investigation in turn will lead to the study of a channel driven by the source supplying information to that channel. It will be shown that the average information for such a system is

$$\text{Expected value of } I(x_k,y_k) = I(X;Y)^* \tag{1-7}$$

[4]

From a physical point of view, the above model may be viewed in a simpler fashion. Consider a source transmitting any one of the two messages x_1 and x_2 with respective probabilities of α and $1 -$. The output of this source is communicated to a receiver via a noisy binary channel. The channel is described by a stochastic matrix:

$$\begin{bmatrix} a & 1-a \\ 1-b & b \end{bmatrix}$$

When x_1 is transmitted, the probability of a correct reception is a and otherwise $1 - a$. Similarly, when x_2 is transmitted, the probability of correct and incorrect receptions are b and $1 - b$, respectively.

It will be shown (Chap. 3) that there is an average amount of information $I(X;Y)$ associated with this model which exhibits the rate of the information transmitted over the channel. This, in turn, raises a basic question. Given such a channel, what is the highest possible rate of transmission of information over

this channel for a specified class of sources? In this manner, one arrives in a natural way at the concept of channel capacity and efficiency of a statistical communication model.

In the above example, the capacity of the channel may be computed by maximizing the information measure $I(X;Y)$ over all permissible values of .

In short, with each probability scheme we associate an entropy which represents, in a way, the average amount of information for the outcomes of the scheme. When a source and a receiver are connected via a channel, several probability schemes such as conditional and joint probabilities have special significance. An important task is to investigate the physical significance and the interrelationships between different entropies in a communication system. The formal treatment of these relations and the concept of channel capacity is presented in several chapters of the text.

The reader acquainted with probability theory may regard information theory as a new branch of that discipline. He can grasp it at a fair speed. The reader without such a background has to move much more slowly. However he will find the introductory material of Chap. 2 of substantial assistance in the study of Chaps. 3 and 4. An introductory treatment of a random variable assuming a continuum of values is given in Chap. 5. Chapter 6 presents a general study of averaging and moments. The reader with such a background will readily recognize the entropy functions that form the nucleus of information theory as moments of an associated logarithmic random variable: $-\log P\{X\}$. Thus the entropy appears to be a new and useful form of moment associated with a probability scheme. This idea will serve as an important link in the integration of information and probability theories. Chapter 7 gives a concise introduction to multinormal distributions, laws of large numbers, and central-limit theorems. These are essential tools for the proof of the main theorems of information theory.

Chapters 8 and 9 extend the information-theory concept to random variables assuming a continuum of values (also continuous signals). The probability background of Chaps. 2, 5, 6, 7, and 10 is in most part indispensable for the study of information theory. However, a few additional topics are included for the sake

of completeness, although they may not be directed toward an immediate applicaton.

Chapter 10 presents a bird's-eye view of stochastic theory, followed by Chap. 11, which studies the information theory of stochastic models. A slightly more advanced consideration (but perhaps the heart of the subject) appears in Chap. 12.

A main application of the theory thus far seems to be in the devising of an efficient matching of the information source and the channel, the so-called *coding theory.* The elements of this theory appear in Chaps. 4 and 13. The Appendix is designed to introduce the reader to a few of the many topics available for further reading in this field.

1-6. Main Contributors to Information Theory.

The historical background of information theory cannot be covered in a few pages. Fortunately there are several sources where the reader can find a historical review of this subject, e.g., The Communication of Information, by E. C. Cherry (*Am. Scientist*, October, 1952), and "On Human Communication," by the same author (John Wiley & Sons, Inc., 1957). (In Chap. 2 of the latter book, Cherry gives a very interesting historical account of developments leading to the discovery of information theory, particularly the impact of the invention of *telecommunication.*)

As far as the communication engineering profession is concerned, it seems that the first attempt to define a measure for "the amount of information" was made by R. V. L. Hartley[5] in a paper called Transmission of Information (*Bell System Tech. J.*, vol. 7, pp. 535–564, 1928).

Hartley suggested that "information" arises from the successive selection of symbols or words from a given vocabulary. From an alphabet of D distinct symbols we can select D^N different words, each word containing N symbols. If these words were all equiprobable and we had to select one of them at random, there would be a quantity of information I associated with such a selection. Furthermore, Hartley suggested the logarithm to the base 10 of the number of possible different words D^N as the quantity of information $I = N \log D$.

The main contributions, which really gave birth to the so-called information theory, came shortly after the Second World War from the mathematicians C. E. Shannon and N. Wiener. Wiener's mathematical contributions to the field of Fourier series and later to time series, plus his genuine interest in the field of communication, led to the foundation of communication theories in general. His two books, "Cybernetics" and "Extrapolation, Interpolation, and Smoothing of Stationary Time Series" (1948 and 1949), paved the way for the arrival of new statistical theories of communication. In a paper entitled The Mathematical Theory of Communication (*Bell System Tech. J.*, vol. 27, 1948), Shannon made the first integrated mathematical attempt to deal with the new concept of the amount of information and its main consequences. Shannon's first paper, along with a second paper, laid the foundation for the new science to be named *information theory.* Shannon's earlier contribution may be summarized as follows:

1. Definition of the amount of information from a semiaxiomatic point of view.

2. Study of the flow of information for discrete messages in channels with and without noise (models of Figs. 1-1 and 1-3).

3. Defining the capacity of a channel, that is, the highest rate of transmission of information for a channel with or without noise.

4. In the light of 1, 2, and 3, Shannon gave some fundamental encoding theorems. These theorems state roughly that for a given source and a given channel one can always devise an encoding procedure leading to the highest possible rate of transmission of information.

5. Study of the flow of information for continuous signals in the presence of noise, as a logical extension of the discrete case.

Subsequent to his earlier work, Shannon has made several additional contributions. These have considerably strengthened the position of the original theory.

Following Wiener's and Shannon's works an unusually large number of scientific papers appeared in the literature in a relatively short time. A bibliography of information theory and allied topics might now, 13 years after

the publication of Shannon's and Wiener's works, contain close to 1,000 papers. This indicates the great interest and enthusiasm (perhaps overenthusiasm) of scientists toward this fascinating new discipline. Here it would be impossible to give a detailed account of the contributions in this field. The reader may refer to A Bibliography of Information Theory, by F. L. Stumpers, and also to *IRE Transactions on Information Theory* (vol. IT-1, no. 3, pp. 31–47, September, 1955).

Even though a historical account has not been attempted here, the names of some of the contributors should be mentioned in passing. Bell Telephone Laboratories appears to be the birthplace of information and coding theory. Among the contributors from Bell Labs are E. N. Gilbert, R. W. Hamming, J. L. Kelley, Jr., B. McMillan, S. 0. Rice, C. E. Shannon, and D. Slepian. P. Elias, R. M. Fano, A. Feinstein, D. Huffman, C. E. Shannon, N. Wiener, and J. A. Wozencraft of the Massachusetts Institute of Technology have greatly contributed to the advancement of information and coding theory. Information theory has received significant stimuli from the works of several Russian mathematicians. A. 1. Khinchin, by employing the results of McMillan and Feinstein, produced one of the first mathematically exact presentations of the theory. Academician A. N. Kolmogorov, a leading man in the field of probability, and his colleagues have made several important contributions. A few of the other Russian contributors are R. L. Dobrushin, D. A. Fadiev, M. A. Gavrilov, I. M. Gel'fand, A. A. Kharkevich, V. A. Kotelnikov,[6] M. Rozenblat-Rot, V. I. Siforov, and I. M. and A. M. Iaglom.

The afore-mentioned names are only a few of a long list of mathematicians and communication scientists who have contributed to information theory. Some other familiar names are D. A. Bell, A. Blanc-Lapierre, L. Brillouin, N. Abramson, D. Gabor, S. Goldman, I. J. Good, N. K. Ignatyev, J. Loeb, B. Mandelbrot, K. A. Meshkovski, W. Meyer-Eppler, F. L. Stumpers, M. P. Schutzenberger, A. Perez, W. Peterson, A. Thomasian, R. R. Varsamov, J. A. Ville, P. M. Woodward.

A list of those actively engaged in the field would be too long to be included here. Reference to some of the current work will be found in the text and in the bibliography at the end of the book.

For a comprehensive list, the reader is referred to existing bibliographies such as those by Stumpers, Green, and Cherry. Recent contributions to information

theory have been aimed at providing more exact proofs for the basic theorems stated by earlier contributors. A state of steady improvement has been prevailing in the literature.

McMillan, Feinstein, and Khinchin have greatly enhanced the elegance of the theory by putting it on a more elaborate mathematical basis and providing proofs for the central theorems as earlier stated by C. E. Shannon. These contributors have confirmed that under very general circumstances, it is possible to transmit information with a high degree of reliability over a noisy channel at a rate as close to the channel capacity as desired.

J. Wolfowitz derived a strong converse of the fundamental theorem of information theory. Among other important theorems, he proved that reliable transmission at a rate higher than the channel capacity is not possible. In the past 2 or 3 years a large number of scientists have become interested in integrating some of the work on encoding theory within the framework of classical mathematics. Reference will be made to their work in Chaps. 13 and 14.

S. Kullback has described the growth of information theory from its statistical roots and emphasized the interrelation between information theory and statistics (Kullback).

The study of time-varying channels has also received considerable attention. Among those who have contributed are C. E. Shannon, R. A. Silverman and S. H. Chang, and V. I. Siforov and his colleagues.

To sum up, the present trend in information theory seems to be as follows: From an engineering point of view, a search for applications of the theory (radar detection, speech, telephone and radio communication, game and decision theory, and particularly implementation of codes) is evident, while the mathematician is still seeking for more rigor in the foundation of the theory and elegance par excellence.

1-7. An Outline of Information Theory.

If we were to make a two–page résumé of information theory for those scientists with a broad background of probability theory, the following could be suggested.

1. The average amount of information conveyed by a discrete random variable Y about another discrete random variable X is suggested by C. E. Shannon.

$$I(X;Y) = \sum_{i=1}^{n} \sum_{k=1}^{m} P\{X = x_i,\ Y = y_k\} \log \frac{P\{X = x_i,\ Y = y_k\}}{P\{X = x_i\} P\{Y = y_k\}}$$

(1-8)

This definition can be generalized to cover not only the case of two or more random variables assuming a continuum of values but also the more general case of random vectors, generalized functions, and stochastic processes (Gel'fand and Iaglom[7]).

2. The channel is specified by $P\{Y = y_k | X = x_i\}$ for all encountered integers i and k. The largest value of the transinformation $I(X;Y)$ obtained over all possible source distribution $P\{X = x_i\}$ is called the *capacity* of the channel [Shannon (I)].

The definition of the channel capacity can be subjected to generalizations similar to those suggested in 1.

3. Let X and Y be two finite sets of alphabets with $x \in X$ and $y \in Y$. The simplest channel is specified by $P\{\ |x \in X\}$. Now consider words of n symbols selected from the X alphabet. These words will be denoted by $u \in U$ and their corresponding received pairs by $v \in V$. This is an nth–order extension of the channel.

4. Given a source $P\{X = x_i\}$, a channel $P\{\ |X = x_i\}$, and their respective nth–order extensions, then to a specified message ensemble U, we may associate a partitioning of the V space such that

$$
\begin{array}{ll}
u_k \rightarrow B_k & k = 1, 2, \ldots, N \\
B_k \cap B_j = \emptyset & \text{for } k \neq j \quad k = 1, 2, \ldots, N \\
P\{B_k|u_k\} \geq 1 - \lambda & k = 1, 2, \ldots, N \\
 & \lambda \text{ a specified positive} \\
 & \text{number usually very small}
\end{array}
$$

This is a decision scheme which in turn specifies a code (N,n,) [A. Feinstein (I)].

5. The central theme of information theory is the following so-called fundamental coding theorem. Given a source, a channel with capacity C, and two constants

$$0 \leq H \leq C \quad 0 < \quad < 1$$

it can be shown that there are an integer $n = g(\quad, H)$ and a code (N, n,) with (N = function of and n) $\geq 2^{nH}$. This is the coding theorem stating the possibility of transmitting information at a rate $H \leq C$ over a noisy channel under specified circumstances.

6. Further elaborate mathematical treatment of the concepts of information theory was presented by B. McMillan, who extended the definition of source and channels from a Markov chain to stationary processes. A proof of the fundamental theorem of 5 as well as a clear understanding of the concepts involved in 5 is due to Feinstein. Khinchin considerably improved the status of the art in general and gave a proof of the fundamental theorem of 5 for the case of stationary processes. A converse of the fundamental theorem of 5 is due to J. Wolfowitz, who also gave sharper estimates than those given in 5. Remaining questions include the search for more general encoding theorems along the lines suggested in 1. A recent step in this direction was taken by C. E. Shannon.[8] The search for engineering applications, particularly low-error probability codes, is ever increasing.

PROBLEMS

1-1. An alphabet consists of four letters A, B, C, D with respective probabilities of transmission $\frac{1}{3}$, $\frac{1}{4}$, $\frac{1}{4}$, $\frac{1}{6}$. Find the average amount of information associated with the transmission of a letter.

1-2. An independent, discrete source transmits letters selected from an alphabet consisting of three letters A, B, and C, with respective probabilities

$$p_A = 0.7 \quad p_B = 0.2 \quad p_C = 0.1$$

(a) Find the entropy per letter.

(b) If consecutive letters are statistically independent and two–symbol words are transmitted, find all the pertinent probabilities for all two–letter words and the entropy of the system of such words.

1-3. Plot the curve $y = -x \log_2 x$ for

$$0 \leq x \leq 1$$

1-4. A pair of dice are thrown. We are told that the sum of the faces is 7. What is the average amount of information contained in this message (that is, the entropy associated with the probability scheme of having the sum of the faces equal to 7)?

1-5. An alphabet consists of six symbols A, B, C, D, E, and F which are transmitted with the probabilities indicated below:

A	0	½
B	01	¼
C	011	$\frac{1}{8}$
D	0111	$\frac{1}{16}$

E	01111	1/32
F	011111	1/32

(*a*) Find the average information content per letter.

(*b*) If the letters are encoded in a binary system as shown above, find $P\{1\}$ and $P\{0\}$ and the entropy of the binary source.

1-6. A bag contains 100 white balls, 50 black balls, and 50 blue balls. Another bag contains 80 white balls, 80 black balls, and 40 blue balls. Determine the average amount of information associated with the experiment of drawing a ball from each bag and predicting its color. The result of which experiment is, on the average, harder to predict?

1-7. There are 12 coins, all of equal weight except one, which may be lighter or heavier. Using information-theory concepts, show that it is possible to determine which coin is the odd one and indicate whether it is lighter or heavier in not more than three weighings with an ordinary balance.

1-8. Solve Prob. 1-7 when the number of coins is N. What is the minimum number of weighings?

1-9. There are seven coins, five of equal weight and the remaining two also of equal weight but lighter than the first five coins. Find the minimum number of weighings necessary to locate these two coins.[9]

PART 1

DISCRETE SCHEMES WITHOUT MEMORY

. . . choose a set of symbols, endow them with certain properties and postulate certain relationships between them. Next, . . . deduce further relationships between them. . . . We can apply this theory *if* we know the "exact physical significance" of the symbols. That is, if we can find objects in nature which possess exactly those properties and inter-relations with which we endowed the symbols. . . . The "pure" mathematician is interested only in the inter-relations between the symbols. . . . The "applied" mathematician always has the problem of deciding what is the exact physical significance of the symbols. *If* this is known, then at any stage in the theory we know the physical significance of our theorems. But the strength of the chain depends on the strength of the weakest link, and on occasion the link of "physical significance" is exceedingly fragile.

J. E. Kerrich, "An Experimental Introduction to the Theory of Probability"
Belgisk Import Co., Copenhagen

CHAPTER 2

BASIC CONCEPTS OF DISCRETE PROBABILITY

2-1. Intuitive Background.

Most of us have some elementary intuitive notions about the laws of probability, and we may set up a game or an experiment to test the validity of these notions. This procedure is much like the so-called classical approach to the theory of probability, which was commonly used by mathematicians up to the 1930s. However, this approach has been subjected to considerable criticism; indeed, the literature on the subject contains many contradictions and controversies in the writings of the major authors. These arise from the intuitive background used and the lack of well-defined formalism and rigor. Thus, the experiment or game is usually defined by assuming certain symmetries and by accepting certain results a priori, such as the idea that certain possible outcomes are equally likely to occur. For example, consider the following problem: Two persons, A and B, play a game of tossing a coin. The coin is thrown twice. If a head appears in at least one of the two throws, A wins. Otherwise, B wins. Intuitively, it seems that the four following possible outcomes are equally probable:

$$(HH), (HT), (TH), (TT)$$

where H denotes head and T denotes tail. A may assume that his chances of winning the game are ¾, since a head occurs in three out of four cases (to his advantage). On the other hand, the following reasoning may also seem logical. If the outcome of the first throw is H, A wins; there is no need to continue the game. Accordingly, only three possibilities need be considered, namely:

(*H*), (*TH*), and (*TT*)

where the first two cases are favorable to *A* and the last one to *B.* In other words, the probability that *A* wins is really $\frac{2}{3}$ instead of ¾. The intuitive approach in this problem thus seems to lead to two different estimates of probability.[10]

The twentieth century has witnessed enormous advances in the rigorous axiomatic treatment of many branches of mathematics. It is true that the axiomatic approach is essentially present in the familiar euclidean geometry and is, in a way, a very old principle. But it was not until the early twentieth century, when the formal and logical structure of mathematics was given serious, systematic study, that its fundamental and profound implications were recognized. Actually, however, the groundwork for the axiomatic treatment was laid by mathematicians such as Peano, Cantor, and Boole during the middle of the nineteenth century. The later efforts of Hilbert, Russell, Whitehead, and others led to a complete reorientation of the basic formulations, bringing mathematics to its present level.

Although consideration of the axiomatic treatment is not our subject here, it may be interesting to point out its general nature. First, a necessary set of symbols is introduced. Then certain inference or operation rules are given for the desired formal manipulation of the symbols, and a proper set of axioms is determined. The formal system thus created must be consistent; that is, the axioms must be independent and noncontradictory. Strictly speaking, the derivation of the theorems is manipulation of symbols without content, using axioms as a starting point and applying the rules of operation. The fundamental nature of a formal system is by no means obvious, and the limitations are even today under very careful study.

A rather new branch of mathematics exists which deals in an axiomatic manner with properties of various abstract spaces and functions defined over these spaces. This is the so-called "measure theory." In the late 1930s and early 1940s attempts were made to put the probability calculus on an axiomatic basis. The work of Kolmogorov, Doob, and many others has contributed greatly toward this aim. Today formal probability theory is an important branch of measure

theory (in a strictly formal sense), although the epistemological meaning of probability itself is subject to philosophical discussion. This latter aspect has been studied by several profound thinkers (von Neumann, Carnap, Russell, Fisher, Neyman, and many others).

Today engineers and research scientists recognize that they must have a working knowledge of the powerful tools of twentieth-century mathematics. Although completely axiomatic and rigorous treatment of this subject is far beyond the scope of this discussion, a classical presentation would be out of date, as it would completely forgo the important modern contributions to the theory. Under these circumstances, it seems that a survey of the modern theory of probability at a nonprofessional level will be a reasonable compromise. Most engineering students are not very familiar with concepts of probability, and it is important that they gain some appreciation of them.

In what follows, some elementary concepts of the theory of sets or so–called "set algebra" must first be introduced. Then these concepts are used to introduce the fundamental definitions of the theory of probability. Such a presentation allows a much wider application of the probability theory than does the older approach, which is inadequate for attacking a large class of modern problems.

2-2. Sets.

The word *set*, in mathematics, is used to denote any collection of objects specified according to a well-defined rule. Each object in a set is called an *element*, a *member*, or a *point* of the set. If x is an element of the set X, this relationship is expressed by

$$x \in X \qquad x \text{ belongs to } X$$

(2-1)

When x is not a member of the set X, this fact is shown by

$$x \notin X \qquad x \text{ does not belong to } X$$

(2-2)

For example, if X is the set of all positive integers, then

$$5 \in X$$
$$\sqrt{2} \in X$$
$$-3 \notin X$$

A set can be specified by either giving all its elements or stating the requirements for the elements belonging to the set. If a, b, c, and d are the only members of a set X, then we may write either

$$X = \{a,b,c,d\}$$

(2-3)

or

$$X = \{x\}$$

(2-4)

In the latter case x designates a general element of X with the understanding that the rule for identifying the members of X is known. For example, if the set X consists of the number of dots on the faces of a die, then we may write

$$X = \{1,2,3,4,5,6\}$$

If the set X consists of all rectangles with an area of 1 square foot we may write $X = \{ x \}$, denoting by x any general rectangle having the specified area.

When every element of a set A is a member of a set B, we say that A is a *subset* of B. This relationship is expressed by either of the forms

$$A \subset B \qquad A \text{ is contained in } B$$

(2-5)

or

$$B \supset A \qquad A \text{ is a subset of } B$$

(2-6)

For example, if A is the set of positive integers and B the set of all rational numbers, then A is a subset of B.

The sets A and B are said to be *equal* if they have exactly the same elements, that is, if and then

$$\begin{aligned} A &\subset B \\ A &\supset B \\ A &= B \end{aligned}$$

(2-7)

For instance, if the set A consists of the roots of the equation

$$(x + 1)(x^2 - 4)(x - 3) = 0$$

and

$$B = \{-2,-1,0,2,3\}$$
$$C = \{x\} \qquad x \text{ being any integer such that } |x| < 4$$

then

$$\left.\begin{array}{l} C \supset A \\ C \supset B \\ A \supset B \\ A \subset B \end{array}\right\} \quad A = B$$

In many instances, when dealing with specific problems, it is most convenient to confine the discussion to objects that belong to a fixed class of elements. This is referred to as a *universal set.* For example, suppose that, in a certain problem dealing with the study of numbers, it may be required to define the set of all integers *I*, or the set of positive numbers *P*, or the set of perfect square integers *S*. All these sets can be looked upon as subsets of the larger set of all real numbers. This latter set may be considered as the universal set *U*, a definition which is useful in dealing with the specific problem under discussion.

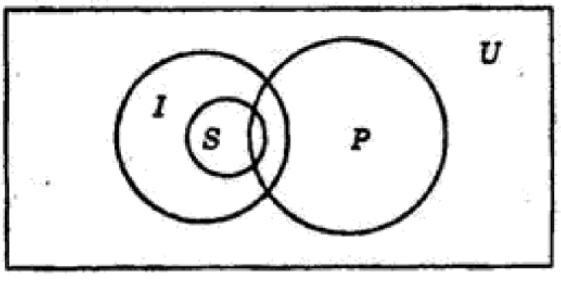


FIG. 2-1. Example of a Venn diagram.

In problems concerned with the interrelationship of sets, an illustrative diagram called a *Venn*[11] *diagram* is of considerable visual assistance.

The elements of the universal set in a Venn diagram are generally shown by points in a rectangle. The elements of any set under consideration are commonly shown by a circle or by any other simple closed contour inside the universal set. The *universe* associated with the aforesaid example is illustrated in Fig. 2-1.

A set may contain a finite or an infinite number of elements. When a set has no element, it is said to be an *empty* or a *null* set. For example, the set of the real roots of the equation is a null set.

$$2z^2 + 1 = 0$$

2-3. Operations on Sets.

Consider a universal set U of any arbitrary elements. U contains all possible elements under consideration. The universal set may contain a number of subsets A, B, C, D, . . . which individually are well-defined sets. The operation of union, intersection, and complement is defined as follows:

The *union* or *sum* of two sets A and B is the set of all those elements that belong to A or B or *both.*

The *intersection* or *product* of two sets A and B is the set of all those elements that belong to both A and B.

The *difference* B–A of any set A relative to the set B is a set consisting of all elements of B that are *not* elements of A.

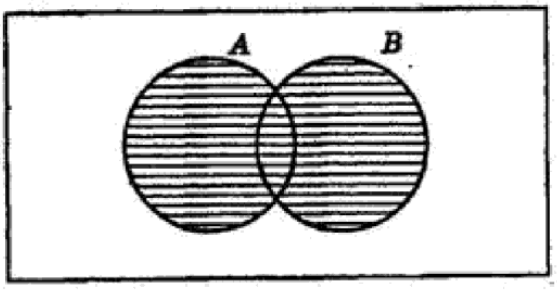


FIG. 2-2. Sum or union A + B.

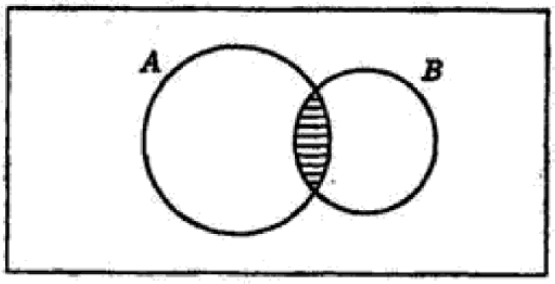


FIG. 2-3. Intersection or product A · B.

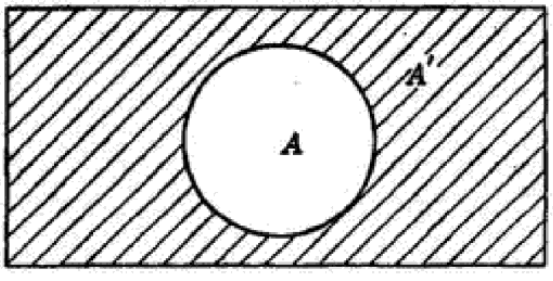


FIG. 2-4. Complement.

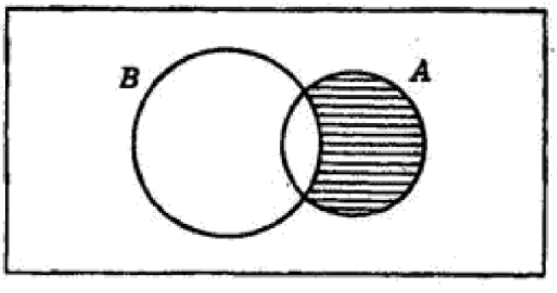


FIG. 2-5. Difference A–B.

The *complement* or *negation* of any set A is the set A' containing all elements of the universe that are *not* elements of A.

In the mathematical literature the following notations are commonly used in

conjunction with the above definitions.

$$A \cup B \qquad A \text{ union } B, \text{ or } A \text{ cup } B$$

(2-8)

$$A \cap B \qquad A \text{ intersection } B, \text{ or } A \text{ cap } B$$

(2-9)

$$A - B \qquad \text{relative complement of } B \text{ in } A$$

(2-10)

$$B \subset A \qquad B \text{ is contained in } A$$

$$\sim A \qquad \text{complement of } A$$

(2-11)

In the engineering literature the notations given below are primarily used.

$$A + B \qquad \text{sum or union}$$

(2-12)

$$A \cdot B \text{ or } AB \qquad \text{intersection or product}$$

(2-13)

$$A - B \qquad \text{difference}$$

(2-14)

$$A' \qquad \text{complement}$$

(2-15)

For the convenience of the engineer we shall generally adhere to the latter notations. However, where any confusion in notation may occur we shall resort to mathematical notation.

The universe and the empty set will be denoted by U and $\emptyset$, respectively. When the product of two sets A and B is an empty set, that is,

$$A \cap B = \emptyset$$

(2-16)

the two sets are said to be *mutually exclusive.* When the product of the two sets A and B is equal to B, then B is a *subset* of A.

$$A \cap B = B \qquad \text{implies} \qquad B \subset A$$

(2-17)

The sum, the product, and the difference of two sets and the complement of any set A are illustrated in the shaded areas of the Venn diagrams of Figs. 2-2 to 2-5. Figures 2-6 and 2-7 illustrate the sets referring to Eqs. (2-16) and (2-17).

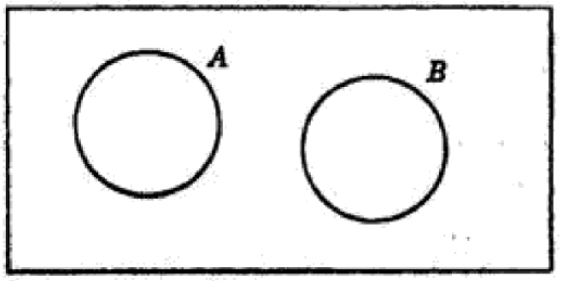


FIG. 2-6. Mutually exclusive sets. AB = 0.

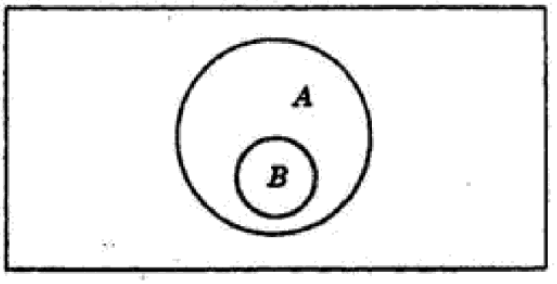


FIG. 2-7. Subset B A. AB = B.

Examples 2-1. Let the universe consist of the set of all positive integers, and let

$$A = \{1,2,3,6,7,10\}$$
$$B = \{3,4,8,10\}$$
$$C = \{x\}$$

where x is any positive integer larger than 5.

Find $A + B$, $A \cdot B$, $A - B$, $A \cdot C$, $B \cdot C$, C , and $A + B + C$.

Solution

$$A + B = \{1,2,3,4,6,7,8,10\}$$
$$A \cdot B = \{3,10\}$$
$$A - B = \{1,2,6,7\}$$
$$A \cdot C = \{6,7,10\}$$
$$B \cdot C = \{8,10\}$$
$$C' = \{1,2,3,4,5\}$$
$$(A + B) + C = U - \{5\}$$

2-4. Algebra of Sets.

We now state certain important properties concerning operations with sets. Let A, B, and C be subsets of a universal set U; then the following laws hold.

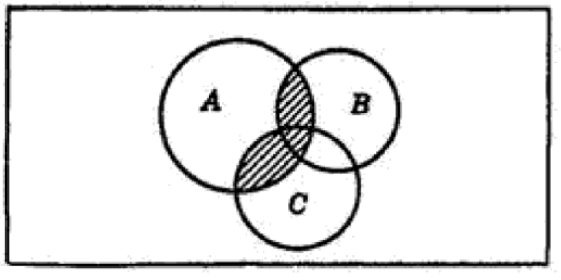


FIG. 2-8. Distributive law. A (B + C) = AB + AC.

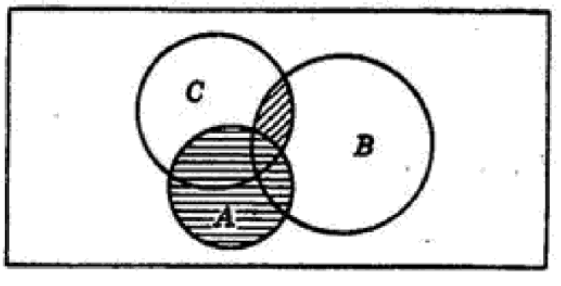


FIG. 2-9. Distributive law. A + BC = (A + B)(A + C).

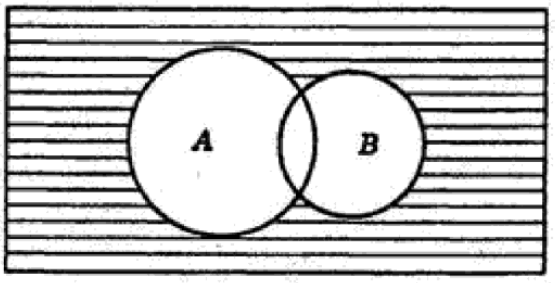


FIG. 2-10. Dualization. (A + B) = A B .

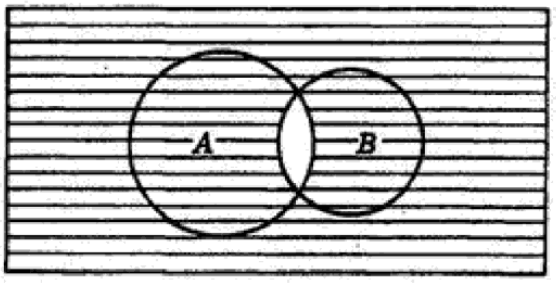


FIG. 2-11. Dualization. (AB) = A + B .

Commutative Laws:

$$A + B = B + A$$
$$AB = BA$$

(2-18)

Associative Laws:

$$(A + B) + C = A + (B + C)$$
$$(AB)C = A(BC)$$

(2-19)

Distributive Laws:

$$A(B + C) = AB + AC$$
$$A + BC = (A + B)(A + C)$$

(2-20)

Complementarity:

$$A + A' = U$$
$$AA' = \emptyset$$

(2-21)

$$A + U = U$$
$$AU = A$$

(2-22)

$$A + \emptyset = A$$
$$A\emptyset = \emptyset$$

(2-23)

Difference Law:

$$(AB) + (A - B) = A$$
$$(AB)(A - B) = \emptyset$$
$$A - B = AB'$$

(2-24)

Dualization or De Morgan's Law:

$$(A + B)' = A'B'$$
$$(AB)' = A' + B'$$

(2-25)

Involution Law:

$$(A')' = A$$

(2-26)

The complement of the set A is the set A.

Idempotent Law: For all sets A,

$$A + A = A$$
$$AA = A$$

(2-27)

While the afore-mentioned laws are not meant to offer an axiomatic presentation of set theory, they are of a fundamental nature for deriving a large variety of identities on sets. The agreement of all these laws with the laws of thought can be verified. One assumes that an element x is a member of the set of the left side of each identity, and then one has to prove that x will necessarily be a member of the set of the right side of the same equation. For instance, in order to prove the distributive law [Eq. (2-20)], let

$$x \quad A(B + C)$$

Then

$$x \in A$$
$$x \in (B + C)$$

Then at least one of the following three cases must be true:

(*a*)

$x \in A$
$x \in B$

(*b*)

$x \in A$
$x \in C$

(*c*)

$x \in A$
$x \in B$
$x \in C$

These are in turn equivalent to

(*a*) $x \in AB$
(*b*) $x \in AC$
(*c*) $x \in ABC$

but $ABC \subset AB$

Therefore it is sufficient to require

$$x \in AB + AC$$

Similarly, one can show that $x \in AB + AC$ implies $x \in A(B + C)$.

The Venn diagram is often a very useful visual aid. Its use is of valuable

assistance in solving problems, as long as the formal proofs are not overlooked.

Example 2-2.. Verify the following relation:

$$(A + B) - AB = AB \quad + A \quad B$$

Solution. By virtue of the third relation of Eqs. (2-24),

$$(A + B) - AB = (A + B)(AB)'$$

Application of De Morgan's law yields

$$(A + B)(AB)' = (A + B)(A' + B')$$
$$(A + B)(A' + B') = A'A + A'B + B'A + B'B = AB' + A'B$$

For an alternative proof, let

$$x \in [(A + B) - AB]$$

Then only one of the following two cases is possible:

(*a*)

$x \in A$

$x \in B$

(*b*)

$x \in B$

$x \in A$

These cases are equivalent to

(*a*) $\left.\begin{matrix} x \in A \\ x \in B' \end{matrix}\right\} x \in AB'$

(*b*) $\left.\begin{matrix} x \in B \\ x \in A' \end{matrix}\right\} x \in A'B$

Note that AB and $A \quad B$ are mutually exclusive sets. Similarly, one can show that all the elements belonging to the set at the right side of the above equation also belong to the set of the left side. Thus the two sides present equivalent sets.

Example 2-3.. Express the set composed of the hatched region of Fig. E2-3 in terms of specified sets.

Solution. The desired set A is

$$A = A_1A_2 + A_2A_3 + A_4A_5A_6$$

See Fig. E2-3.

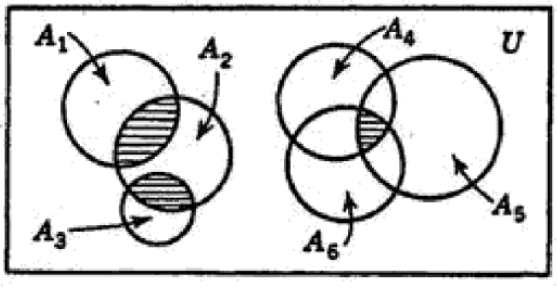


FIG. E2-3

Example 2-4. Verify the relation

$$(A + B) \quad C = C - C(A + B)$$

Solution. We may wish to verify the validity of this relation by using the Venn

diagram of Fig. E2-4.. The left side of this equation represents the part of the set C that is not in A or B. The right side represents $C - CA - CB$, that is, the part of C that is not included either in A or in B.

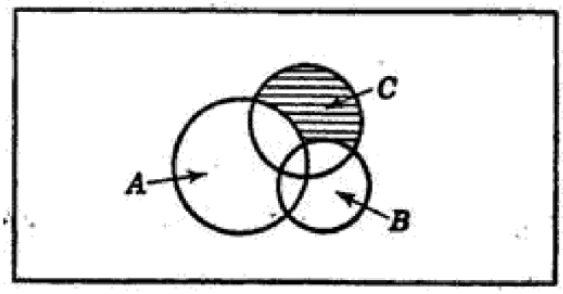


FIG. E2-4

Example 2-5. Consider the relay circuit of Fig. E2-5.. The setup contains coils which must be activated for closing or opening the corresponding relay. A, B, and C are normally open relays and A , B , and C are normally closed relays which are respectively activated by the same controlling source. For instance, when relay A is open because of the effect of its activating coil, A is closed. In order to have a current flow between the terminals M and N, we must have the set of relay operations indicated by $ABC + AB\ C + A\ B\ C$. With this in mind, the question is to replace the given network by a less complex *equiualent* circuit.

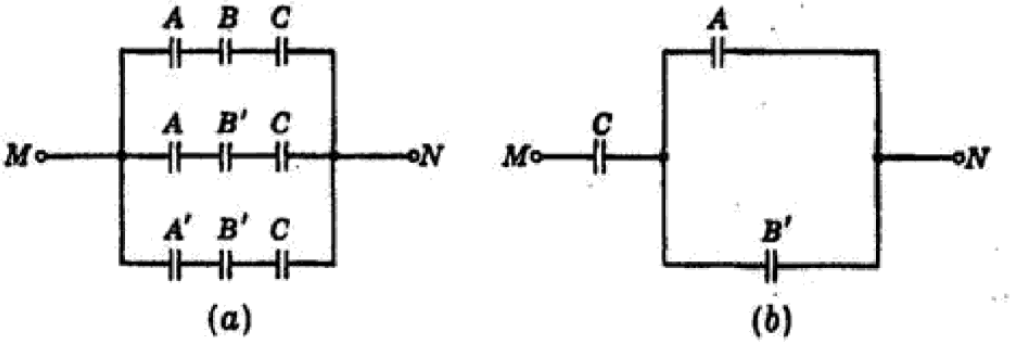


FIG. E2-5

Solution. A way of simplifying the above expression is the following:

$$F = C(AB + AB' + A'B')$$
$$F = C[A(B + B') + A'B']$$
$$F = C(AU + A'B')$$
$$F = C(A + A'B')$$
$$F = C(A + B')$$

A circuit presentation of this example is illustrated in Fig. E2-5*b*.

Example 2-6. Verify the equivalence of the two relay circuits of Fig. E2-6.

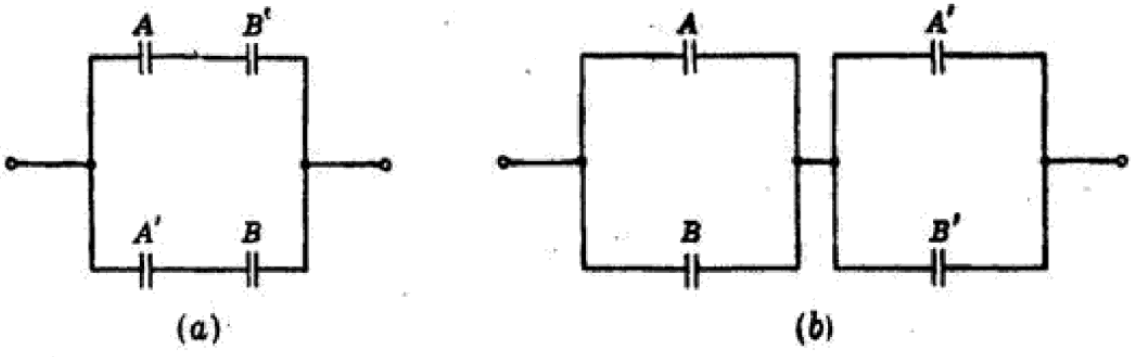


FIG. E2-6

Solution. The set that corresponds to the operation of the circuit in Fig. E2-6*b* is

$$(A + B)(A \quad + B \quad)$$

Direct multiplication gives

$$AA \quad + AB \quad + BA \quad + BB \quad = AB \quad + A \quad B$$

The latter set can be immediately identified with the circuit of Fig. E2-6*b*.

Sheffer-stroke Operation. Examples 2-5 and 2-6 have illustrated some use of Boolean algebra in relay circuits. As another example of the use of Boolean algebra in engineering problems, we discuss briefly what is referred to as the *Sheffer-stroke operation.* This operation for two sets X and Y is denoted by $(X|Y)$ and is defined by the equation

$$(X|Y) = X \cup Y \quad \text{not } X, \text{ or not } Y, \text{ or not } X \text{ and } Y$$

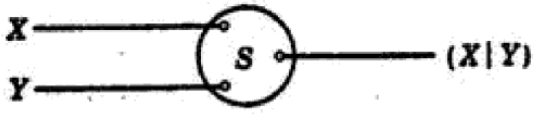


FIG. 2-12.. Sheffer stroke.

The Sheffer stroke commonly illustrated by the three-port diagram of Fig. 2-12 has the distinct property that it can replace all three basic operations of Boolean algebra (sum, product, and negation). The validity of this statement can be exhibited in a direct manner.

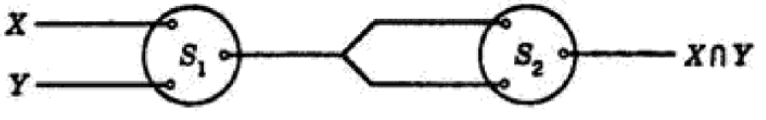


FIG. 2-13.. Product operation by two Sheffer strokes.

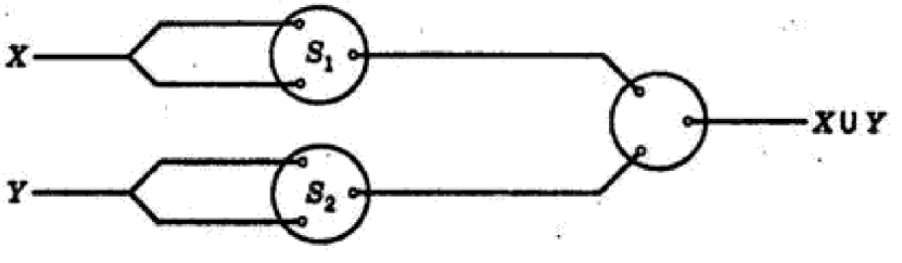


FIG. 2-14.. Summing operation by three Sheffer strokes.

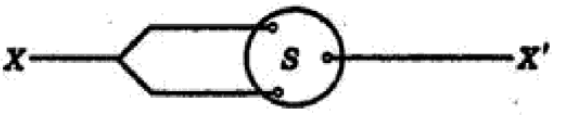


FIG. 2-15.. Operation of negation with a Sheffer stroke.

PRODUCT OPERATION. Reference to the diagram of Fig. 2-13 suggests that

$$\begin{aligned}((X|Y)|(X|Y)) &= (X' \cup Y')' \\ &= ((X \cap Y)')' = X \cap Y\end{aligned}$$

SUMMING OPERATION. The diagram of Fig. 2-14 suggests

$$\begin{aligned}((X|X)|(Y|Y)) &= (X|X)' \cup (Y|Y)' \\ &= X \cup Y\end{aligned}$$

NEGATION. Reference is made to the diagram of Fig. 2-15..

$$(X|X) = X \quad \cup X \quad = X$$

2-5. Functions.

In this section, some well-defined objects or numbers will be associated with each and every element of a given set. The rule on which this relationship is based is commonly known as *function.*

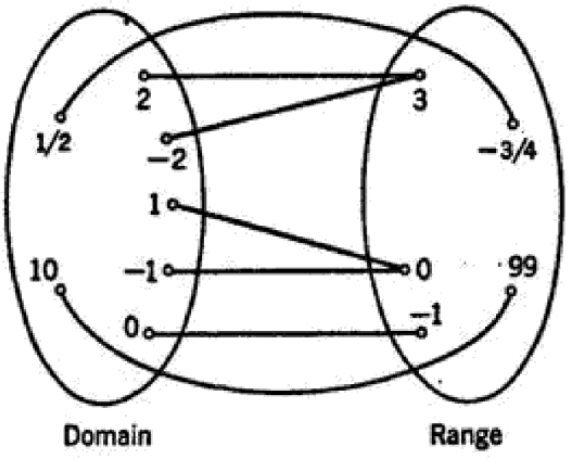


FIG. 2-16.. Domain X and range Y.

If $X = \{x\}$ is a set and $y = f(x)$ is a rule, that is, a sequence of specified operations and correspondence for assigning a well-defined object y to every member of X, then by applying this rule to the set X, we obtain a set $Y = \{y\}$. The set X is called the *domain* and Y the *range.* When x covers the elements of X, then y will correspondingly cover the elements of Y. For example, let X be the set of all persons living in the state of California on January 1, 1959, and let the function be defined as follows: anyone who is the father of a person described by X and is in the state of Colorado on January 1, 1959. Assuming that all the words appearing in the rule, such as father, California, Colorado, are well-defined words, this may be considered as a well-defined function. To each member of X there corresponds an object in the set Y. In this example, element zero in Y corresponds to some of the elements of X, and several members of X might have a unique correspondent in Y.

As another simple example, consider the set

$$X = \{1,2,0,-2,-1,\tfrac{1}{2},10\}$$

and the function

$$f(x) = x^2 - 1$$

which lead to the set

$$Y = \{0,3,-1,3,0,-\tfrac{3}{4},99\}$$

The domain of x and the range of y are shown in Fig. 2-16, the correspondence being one-to-one from X to the Y set.

Example 2-7. A set of ordered pairs $s = \{(X,Y)\}$, that is, a set of points in the rectangular coordinate system, is given in Fig. E2-7*a*.

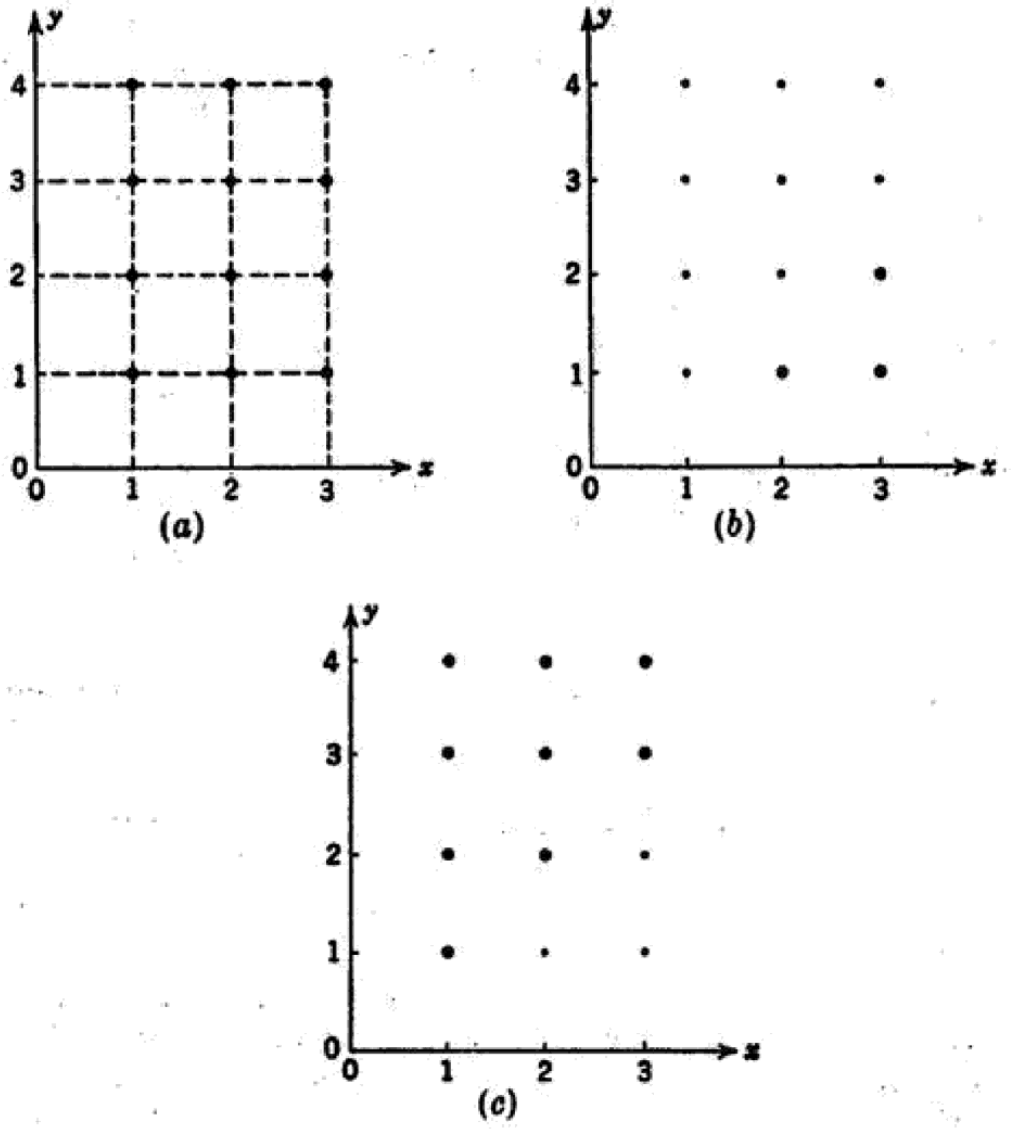


FIG. E2-7

(*a*) Describe the elements of the subset $a = \{(X,Y) \mid y < x\}$.

(*b*) Describe the elements of the subset $b = s - a$. *Solution*

(*a*) See Fig. E2-7*b*.

(*b*) See Fig. E2-7*c*.

Numerical Functions. Functions that have numerical values are the most

common type. We can define the basic algebraic operations for a family of numerical functions defined over a specific domain $X = \{x\}$. For instance, if $f_1(x)$, $f_2(x)$, and $f_3(x) = \text{const} = k$ have a common domain,

$$\begin{gathered} f_1(x) + f_2(x) \\ kf_1(x),\ kf_2(x) \\ f_1(x) \cdot f_2(x) \end{gathered}$$

(2-28)

are also defined over the same domain.

As a particularly interesting case of numerical function, consider the correspondence between the elements of a set having a finite number of elements and a set of positive integers. Such functions have the following basic property: If A and B are two disjoint sets having a number of a and b elements, respectively, then the number of elements of the set $A + B$ is

$$n(A \cup B) = n(A) + n(B) = a + b$$

(2-29)

where $n(X)$ means the number of elements in the set X. The number of elements of a finite set has the simple but important property of being a real additive function. In other words, assume that A and B are themselves subsets of a set S containing a finite number of subsets $A, B, C, D, \ldots$. Let f be a function that assigns a real number $f(X)$ to each X S, such that for any two disjoint subsets of S we have

$$f(A \cup B) = f(A) + f(B)$$

Then f is called an *additive set function.* This result, of course, holds for the union of a finite number of disjoint subsets of S.

Equivalent Sets. Let A and B be two sets. A rule that associates with each element $a \in A$ exactly an element $b \in B$, and conversely, is said to be a one-to-one correspondence between A and B. Two sets A and B are equivalent if, and only if, a one-to-one correspondence between their elements can be established.

As an example, consider the set of all persons (A) living in New York State and (B) living in the state of Arizona at a given time. Now if we associate each person of A with the cardinal numbers 1 to N, inclusive, and each person of B with the cardinal numbers 1 to M, inclusive, it is clear that there is a one-to-one correspondence between the elements of $A + B$ and the set of cardinal numbers 1 to $M + N$, inclusive.

The number of elements in a set may or may not be finite. In the latter case, if the elements of the set can be placed in a one-to-one correspondence with the set of natural numbers

$$\{1,2,3, \ldots\}$$

(2-30)

we say that the set has a *denumerable* or countable number of elements. For example, the number of elements in the set

$$\{1,4,9,16, \ldots, n^2, \ldots\}$$

is denumerably infinite.

A common example of nondenumerable sets can be given by considering points on a straight line. Let x denote the abscissa of a point of the line segment between points A and B with respective abscissa a and b. The inequality

$$a < x < b$$

indicates a set of points on the line AB that does not contain the end points A and B. Such a set is termed an *open interval* and is denoted by

$$]a,b[\qquad \text{open interval} \quad a < x < b$$

(2-31)

Similarly, a *closed interval* is defined and denoted as follows:

$$[a,b] \qquad \text{closed interval} \quad a \leq x \leq b$$

(2-32)

It can be shown that the number of points in [0,1] are nondenumerable.[12] If the set A is equivalent to the set of points in [0,1], it is said that A has the power of continuum. [13]

The additive property of the function under consideration, i.e., the number of elements in finite sets, makes the following relations self-evident.

$$n(A \cup B) = n(A) + n(B) - n(AB)$$
$$n(A - B) = n(A) - n(AB)$$

(2-33)

$$n(A) + n(A') = n(U)$$

(2-34)

For a set containing three subsets A, B, and C one can derive

$$n(A \cup B \cup C) = n[(A) \cup (B \cup C)]$$
$$n(A \cup B \cup C) = n(A) + n(B \cup C) - n(AB \cup AC)$$
$$n(A \cup B \cup C) = n(A) + n(B) + n(C) - n(BC) - n(AB) - n(AC) + n(ABC)$$

(2-35)

The following example is designed to employ the additive property of the afore-mentioned set functions.

Example 2-8. There are three radio stations A, B, and C which can be received in a town of 3,000 families. The following data are given:

(a)1,800 families listen to station A.
(b)1,700 families listen to station B.
(c)1,200 families listen to station C.
(d)1,250 families listen to stations A and B.
(e)700 families listen to stations A and C.
(f) 600 families listen to stations B and C.
(g)200 families listen to stations A, B, and C.

Of course any family may listen to other stations besides the ones specified in each case. The problem is to obtain the number of families who are not listening to any station.

Solution. We draw the pertinent Venn diagram of Fig. E2-8 and, starting from the bottom of the above list, indicate the corresponding number of elements of each subset on the diagram. The number of families in set g is 200. Thus, the number of families listening to B and C but not to A is

$$n(BCA\quad) = n(BC) - n(BCA) = 600 - 200 = 400$$

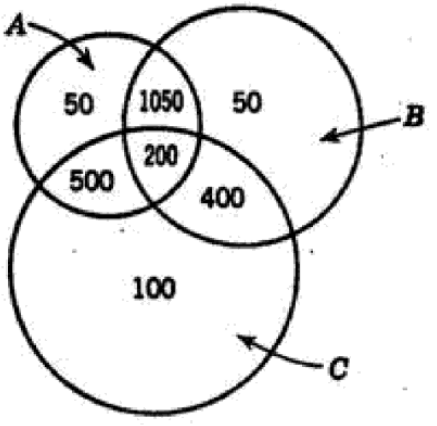


FIG. E2-8

Following this procedure one can obtain all the numbers associated with each disjoint set in the Venn diagram. The total number of families listening to one or more stations is 2,350. This indicates that there are 650 families not listening to any of the above radio stations.

Similar questions can easily be answered by referring to the Venn diagram of Fig. E2-8. For example, the number of families who are not listening to A but are listening either to B or to C or to both is

$$n\{A'(B \cup C)\} = n(A'B) + n(A'C) - n(A'BC)$$
$$n\{A'(B \cup C)\} = 450 + 500 - 400 = 550$$

2-6. Sample Space.

In this section we shall make preparations for applying the concept of set theory to probability. When talking about probability we usually have in mind what can be termed an *experiment* with certain *outcomes.* An outcome is any one of the possibilities that may be expected from the experiment. The totality of all these outcomes forms a universal set which is called the *sample space.* Each outcome is a *point* of the sample space.

For example, the throw of an ordinary die may be considered as an experiment having six possible outcomes. With this experiment we associate a

universal set containing six points, each corresponding to one of the outcomes of the experiment:

$$\{1,2,3,4,5,6\}$$

If the die is thrown twice, the sample space associated with the experiment contains 36 points corresponding to the following outcomes:

11	12	13	14	15	16
21	22	23	24	25	26
31	32	33	34	35	36
41	42	43	44	45	46
51	52	53	54	55	56
61	62	63	64	65	66

A sample space may be finite or infinite, if it contains a finite or an infinite number of points, respectively. The sample space corresponding to a single throw of a die is finite. On the other hand, the sample space corresponding to an experiment of throwing the die until a 6 appears is an infinite space. It is possible to conceive a situation where one may have to throw the die infinitely many times without obtaining a 6. A sample space containing at most a denumerable number of elements is termed *discrete.* Sample spaces containing a nondenumerable number of elements include the so-called "continuous sample space." In this case the range of the elements covers a continuum of values in contrast with the discrete set of values in the discrete sample space.

A subset of a sample space is called an *event.* Thus, an event is a subset of a sample space containing any number of points or outcomes.

(See Fig. 2-17.)

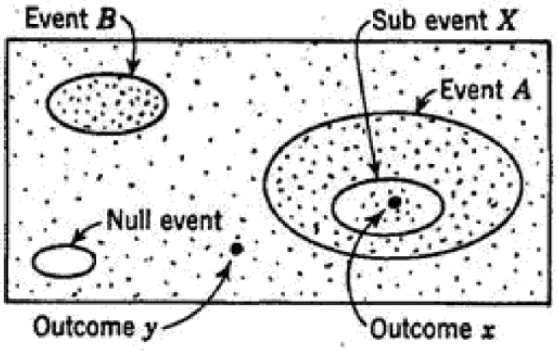


FIG. 2-17. A probability space.

An event containing no outcomes is a null set or an empty set and represents an event that is impossible. An event containing all sample points is an event that is certain to occur. This may be denoted by the universal set U, which means that the event under consideration is bound to occur. The outcome of an event implies the occurrence of any one of its possible outcomes. The following glossary of terms may be of assistance in the transition from the language of set theory to that of probability theory:

U	All possibilities.
$A \quad U$	A particular event.
$A = U$	The event A must occur (certain).
$A = \emptyset$	The event A is impossible.
A'	The event A does not occur.
$x \in X$ or $X \subset A$	x is any particular outcome of X. The occurrence of x implies the occurrence of A and X.
$y \in A$	y is not an outcome of the event A.
$ABC \bullet\bullet\bullet D = S$	S is the event of the simultaneous occurrence of events $A,B,C, \ldots D$.
$A + B + C + \bullet\bullet\bullet + D = S$	S is the event of the occurrence of A or B or C

	or . . . or D, or any combination of these.
$ABC \bullet\bullet\bullet D = \emptyset$	The events $A,B,C, \ldots, D$ are incompatible.
$A+B+C+ \bullet\bullet\bullet +D = U$	At least one of the events $A, B, C, \ldots, D$ must occur.

Example 2-9. A traveler has the choice of traveling by car, train, plane, or any combination of the three for a particular trip. Define the sample space and express some of the events of interest.

Solution. Let C, T, and P correspond to the fact of traveling by car, train, or plane, respectively. The following events are self-explanatory.

CTP	traveling by car, train, and plane
CTP	traveling by car and train but not by plane
CT	traveling by car and train (with or without plane)
$C + T$	traveling by car, by train, or by car and train (may or may not take the plane)
$U - P$	not traveling by plane

Example 2-10. A traveler travels between cities M and N. The possible roads are shown in Fig. E2-10. Define the sample space and the events that the traveler goes through towns A, B, or both.

Solution. Assuming that the traveler does not change direction while traveling, the following selection of roads is possible:

15, 25, 146, 147, 246, 247, 36, 37, 345

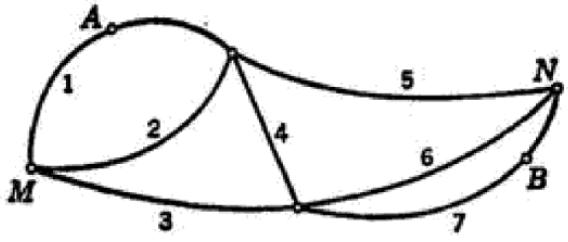


FIG. E2-10

The sample space has nine points; i.e., our defined experiment may have nine distinct outcomes. The event of passing through the town A (event E_1) consists of any of the three points 15, 146, and 147. The event of passing through the town B (event E_2) consists of any of the three points 147, 247, and 37. Finally the event of passing through A and B (event E_1E_2) consists of a single point 147. Similarly, the following events can easily be identified:

E_1E_2'	15, 146
$E_1'E_2$	37, 247
$E_1 + E_2$	15, 146, 147, 247, 37
$E_1'E_2'$	25, 246, 36, 345

2-7. Probability Measure.

In Sec. 2-5 on Functions we have associated arbitrary set functions with the elements of sets. In particular, we have outlined some numerical functions and observed certain rules such as the additivity relation of Eq. (2-33), when the set function was the number of elements in each set. The study of the mathematics of set functions has its place in a branch of mathematics known as *measure theory.* The *probability measure* is a specific type of function which can be associated with sets. When dealing with abstract mathematics, one may specify any arbitrary properties for the measure. However, the end result in this study is probability as applied to the physical world. It is mainly for this reason that we require our probability measure to fulfill the requirements that will be described below. These requirements are matters of convenience for our subsequent

dealing with physical problems rather than a mathematical necessity.

An experiment is defined so that to each possible outcome of this experiment there corresponds a point in the sample space. The number of outcomes of this experiment is assumed to be at most denumerable. The outcomes are labeled by symbols a_k, and a single-valued real function $m\{a_k\}$ called the *probability measure* is defined. An event of interest A is considered as the set of the outcomes a_k giving rise to that event. The probability measure of an event is defined as the sum of the probability measures associated with all the outcomes a_k of that event. Two events A and B are termed *disjoint* if they contain no outcome in common. That is, two disjoint events cannot happen simultaneously. The probability measure has the following assumed properties:

$$0 \leq m\{A_k\}$$

(2-36) [14]

$$m\{A \cup B\} = m\{A\} + m\{B\} \qquad \text{if } A \text{ and } B \text{ are disjoint}$$

(2-37)

$$m\{X\} = 0 \qquad \text{if } X = \emptyset$$

(2-38)

$$m\{X\} = 1 \qquad \text{if } X = U$$

(2-39)

For a more general case involving a continuous sample space one employs the concept of integration. This is not considered here as it requires rigorous mathematical treatment beyond the present scope of interest. The interested

reader is referred to "Probability Theory" by M. Loève (Chap. 1).

The above measure-theory approach is certainly valid. Any measure satisfying the specified requirements, when applied to a problem involving sets, will lead to a consistent mathematical setup. For example, if A, B, and C are subsets of a universal set U with an additive probability measure, that is, the measure associated with the union of two disjoint sets is equal to the sum of their individual measures, the following relations are valid:

$$\left.\begin{aligned} m\{A\} &\leq m\{B\} \\ m\{A\} &= m\{B\} - m\{B - A\} \end{aligned}\right\} \quad \text{if } A \subset B$$

(2-40)

(2-40a)

$$m\{A'\} = m\{U - A\} = m\{U\} - m\{A\} = 1 - m\{A\}$$

(2-41)

$$m\{A \cup B\} = m\{(A - AB) \cup B\} = m\{A\} - m\{AB\} + m\{B\}$$

(2-42)

$$m\{A\} + m\{B\} \geq m\{AB\}$$

(2-43)

For three disjoint sets,

$$m\{A \cup B \cup C\} = m\{A\} + m\{B\} + m\{C\}$$

(2-44)

For three sets in general,

$$m\{A \cup B \cup C\} = m\{A\} + m\{B\} + m\{C\} - m\{AB\} - m\{BC\} - m\{CA\} + m\{ABC\}$$

(2-45)

Example 2-11. Consider a set of all intervals I contained in the closed interval [0,1]. With each and every interval I we associate a measure function $L(I)$ equal to the ordinary length of the same interval. See if such a measure satisfies the requirements of a probability measure.

Solution. The requirement of (2-36) is satisfied, as the length associated with each member of the set is a nonnegative number between 0 and 1. The condition (2-37) is fulfilled by nonoverlapping intervals (mutually exclusive sets). The requirements (2-38) and (2-39) are also met. For a more thorough discussion the reader is referred to Creamèr (Chap. 4, The Lebesgue Measure of a Linear Point Set).

2-8. Frequency of Events.[15]

In Sec. 2-7 on Probability Measure an introductory axiomatic account of probability as a measure of a set was given. The object of this section is to supplement the set-theory point of view with some perhaps less formal discussion of the probability of occurrence of certain events of a defined experiment. In other words, we wish to make a transition from the suggested abstract mathematical measure to some empirical numerical function fulfilling the specified measure requirements.

The first step toward this objective is to define an experiment such as the tossing of a coin or the drawing of a card from a given deck of cards. Next, all the outcomes of this experiment must be specified. Now consider a specific event X_k among all the possible events of the experiment under consideration. If the basic experiment is repeated N times among which the event X_k has appeared $n(X_k)$ times, the ratio

$$\frac{n(X_k)}{N}$$

is defined as the *relative frequency* of the occurrence of the event X_k. In case N is increased indefinitely, intuitively speaking, the "limit" of

$$\frac{n(X_k)}{N}$$

(2-46)

as $N \to \infty$ is the probability $P\{X_k\}$ of the event X_k. This "definition" of probability is more elaborate than the classical definition of Laplace which defines the probability as the ratio of the number of favorable events to the total number of possible events. In the latter definition all events are considered to be equally likely, that is, throwing of a true die by an honest person under prescribed circumstances. It is to be noted that

$$0 \leq n(X_k) \leq N$$

(2-47)

$$0 \leq \frac{n(X_k)}{N} \leq 1$$

(2-48)

$$0 \leq \lim_{N \to \infty} \frac{n(X_k)}{N} \leq 1$$

(2-49)

Equation (2-49) states that the probability of any event X_k is a real number in the real interval [0,1].

$$0 \leq P\{X_k\} \leq 1$$

(2-50)

Considering an event that occurs in every observation yields the limiting case $P\{X\} = 1$, which is a *certain event.* Also, an event that never occurs will lead to the other limiting case, $P\{X\} = 0$, which is an *impossible event.*

We have thus far shown that this empirical definition of probability satisfies the requirements (2-36), , (2-38), and (2-39). It remains to be seen whether the requirement (2-37) holds or not. In order to verify this, consider two particular events A and B among the events that result from the experiment. Let the experiment be repeated n times. Each observation can belong to only one of the four following categories (Fig. 2-18) :

1. A has occurred but not B, the event AB .
2. B has occurred but not A, the event BA .
3. Both A and B have occurred, the event AB.
4. Neither A nor B has occurred, the event A B .

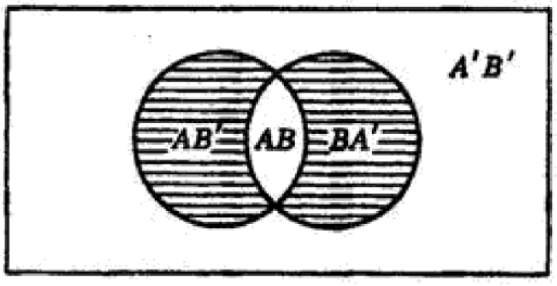


FIG. 2-18.. Probability space of two events.

Note that

$$A = AB' \cup AB$$
$$B = BA' \cup AB$$
$$A \cup B = AB' \cup AB \cup BA'$$

If the number of events of each category is denoted by n_1, n_2, n_3, and n_4, respectively, the following equations are self-explanatory:

$$n_1 + n_2 + n_3 + n_4 = n$$

(2-51)

$$f\{A\}, \text{ relative frequency of } A \text{ independent of } B = \frac{n_1 + n_3}{n}$$

(2-52)

$$f\{B\}, \text{ relative frequency of } B \text{ independent of } A = \frac{n_2 + n_3}{n}$$

(2-53)

$$f\{A + B\}, \text{ relative frequency of either } A, B, \text{ or both} = \frac{n_1 + n_2 + n_3}{n}$$

(2-54)

$$f\{AB\}, \text{ relative frequency of } A \text{ and } B \text{ occurring together} = \frac{n_3}{n}$$

(2-55)

$f\{A|B\}$, relative frequency of A under condition that B has occurred

$$= \frac{n_3}{n_2 + n_3}$$

(2-56)

$f\{B|A\}$, relative frequency of B under condition that A has occurred

$$= \frac{n_3}{n_1 + n_3}$$

(2-57)

When the number of experiments tends to infinity, these simple relations with proper interpretation lead to the addition law and multiplication law:

$$P\{A \cup B\} = P\{A\} + P\{B\} - P\{AB\}$$

(2-58)

$$P\{AB\} = P\{A\}P\{B|A\}$$

(2-59)

$$P\{AB\} = P\{B\}P\{A|B\}$$

(2-60)

For the special case of mutually exclusive events, $P(AB) = 0$,

$$P\{A + B\} = P\{A\} + P\{B\}$$

(2-61)

Equation (2-61) shows the validity of the requirement (2-37) for the chosen set function termed the relative frequency of the event.

More specifically, we have proved that the probability measure defined by Eq. (2-46) satisfies the following basic properties for all sets defined in sample space:

$$0 \leq P\{A\} \leq 1$$

(2-62)

$$P\{A \cup B\} = P\{A\} + P\{B\} \qquad \text{for mutually exclusive } A \text{ and } B$$

(2-63)

$$P\{X\} = 0 \qquad \text{if, and only if, } X = \emptyset$$

(2-64)

$$P\{X\} = 1 \qquad \text{if, and only if, } X = U$$

(2-65)

Therefore the suggested definition of the frequency can serve as a probability measure. The implication of Eqs. (2-58) to (2-60) will be investigated in subsequent sections.

The frequency approach is a rather common approach for defining the probability when dealing with physical problems. Its mathematical concept relies on the tacit assumption of an equiprobable measure, that is, the equal likelihood of the outcome of the repeated experiments. We assume that the measure associated with an event, in the case of the repeated experiment, is

proportional to the number of the outcomes in the event under consideration. In essence, this assumption makes the frequency definition somewhat too restrictive.

2-9. Theorem of Addition.

It seems appropriate now to continue with our formalism without restriction to an immediately practical but slow procedure. For two events A and B of the sample space one has

$$A \cup (B - AB) = A \cup B$$

(2-66)

The additive property of the probability measure in Sec. 2-7 suggests that

$$\begin{aligned} m\{A + B\} &= m\{A\} + m\{B\} - m\{AB\} \\ P\{A + B\} &= P\{A\} + P\{B\} - P\{AB\} \leq P\{A\} + P\{B\} \end{aligned}$$

(2-67)

If two events A and B are mutually exclusive, then

$$P\{AB\} = P\{\emptyset\} = 0$$

(2-68)

$$P\{A \cup B\} = P\{A\} + P\{B\}$$

(2-69)

For two opposite events A and A , one has $A + A$ = U, and since AA' = 0, then

$$P\{A \cup A'\} = P\{A\} + P\{A'\} = P\{U\} = 1$$
$$P\{A'\} = 1 - P\{A\}$$

(2-70)

For the three events A, B, and C, we may write

$$P\{A \cup B \cup C\} = P\{A \cup B\} + P\{C\} - P\{(A \cup B)C\}$$

(2-71)

$$P\{A \cup B \cup C\} = P\{A\} + P\{B\} + P\{C\} - P\{AB\} - P\{BC\} - P\{CA\} + P\{ABC\}$$

(2-72)

This is indeed made clear by employing a pertinent Venn's diagram, $P\{ABC\}$ being the probability of the simultaneous occurrence of the three events. If the events are mutually exclusive, then

$$P\{A \cup B \cup C\} = P\{A\} + P\{B\} + P\{C\}$$

(2-73)

More generally, for a number of events $A_1, A_2, \ldots, A_n$ one may write

$$P\{A_1 \cup A_2 \cup \cdots \cup A_n\} = P\{A_1\} + P\{A_2\} + \cdots + P\{A_n\} - P\{A_1A_2\} - P\{A_1A_3\} - \cdots - P\{A_{n-1}A_n\} + P\{A_1A_2A_3\} + P\{A_1A_2A_4\} + \cdots + P\{A_{n-2}A_{n-1}A_n\} + \cdots + (-1)^{n-1}P\{A_1A_2 \cdots A_n\} \quad (2-$$

(2-74)

By extension of the relation in Eq. (2-66), it can be shown without difficulty that

$$P\{A_1 \cup A_2 \cdots A_n\} \leq P\{A_1\} + P\{A_2\} + \cdots + P\{A_n\}$$

(2-75)

The equality sign holds when the events A_k and A_j are mutually exclusive for all $k \neq j$.

Example 2-12. An urn contains 11 balls numbered from 1 to 11. If a ball is selected at random, what is the probability of having a ball with a number which is a multiple of either 2 or 3?

Solution. Let A and B be the events that the ball number is a multiple of 2 and 3, respectively. The event of interest is $A + B$.

$$\begin{aligned} P\{A\} &= 5/11 \\ P\{B\} &= 3/11 \\ P\{AB\} &= 1/11 \\ P\{A + B\} &= 5/11 + 3/11 - 1/11 = 7/11 \end{aligned}$$

Example 2-13. One card is drawn from a regular deck of 52 cards. What is the probability of the card being either red or a king?

Solution. Let A be the event that the card is red, and B the event that the card is a king. The event of interest is $A + B$. Where A and B are not exclusive events, apply Eq. (2-67):

$$\begin{aligned} P\{A\} &= 1/2 \\ P\{B\} &= 1/13 \\ P\{AB\} &= (1/13)(1/2) = 1/26 \\ P\{A + B\} &= 1/2 + 1/13 - 1/26 = 7/13 \end{aligned}$$

Example 2-14. An honest coin is tossed 10 times. What is the probability of having at least (*a*) one tail and (*b*) two tails?

Solution. The main assumption in this and in similar problems is the concept of independence of successive trials and the equally probable outcomes.

Let A and B be the events of getting no tail and exactly one tail, respectively. Then

$$P\{A\} = \left(\frac{1}{2}\right)^{10} = \frac{1}{1,024}$$
$$P\{B\} = 10\left(\frac{1}{2}\right)^{10} = \frac{10}{1,024}$$

The events of interest are

$$U - A = A'$$

(*a*) $$P\{A'\} = 1 - \frac{1}{1,024} = \frac{1,023}{1,024}$$

$$U - (A + B) = (U - A) - B = A' - B$$

(*b*) $$P\{A' - B\} = \frac{1,023}{1,024} - \frac{10}{1,024} = \frac{1,013}{1,024}$$

2-10. Conditional Probability.

Consider two events A and B. The conditional probability of event A based on the hypothesis that event B has occurred is defined by the following relation:

$$P\{A|B\} = \frac{P\{AB\}}{P\{B\}} \qquad P\{B\} \neq 0$$

(2-76)

The use of this definition can be justified by returning to the previously treated case of Sec. 2-8. The frequency of the occurrence of event A under the assumption that B has occurred is

$$f\{A|B\} = \frac{n_3}{n_2 + n_3} = \frac{f\{AB\}}{f\{B\}}$$

(2-77)

By the same token, the frequency of the occurrence of B, knowing that A has already occurred, is

$$f\{B|A\} = \frac{n_3}{n_1 + n_3} = \frac{f\{AB\}}{f\{A\}}$$

(2-78)

Increasing the number of trials indefinitely gives

$$P\{A|B\} = \frac{P\{AB\}}{P\{B\}} \qquad P\{B\} \neq 0$$

(2-79)

$$P\{B|A\} = \frac{P\{AB\}}{P\{A\}} \qquad P\{A\} \neq 0$$

(2-80)

The two events A and B are said to be *mutually independent* if

$$P\{A|B\} = P\{A\}$$
$$P\{B|A\} = P\{B\}$$

(2-81)

Note that for mutually independent events

$$P\{AB\} = P\{A\} \cdot P\{B\} \tag{2-82}$$

Equations (2-81) and (2-82) are alternatively used as the defining relations for two mutually independent events.[16]

Example 2-15. Three boxes of identical appearance contain two coins each. In one box both are gold, in one box both silver, and in the third box one is a silver coin and the other is a gold coin. Suppose that a box is selected at random and, further, that a coin in that box is selected at random. If this coin proves to be gold, what is the probability that the other coin in the box is also gold?

Solution. Let

A_{gg} be the event that the other coin in the selected box is also gold (that is, the selected box is *gg* box)

B_g be the event that the first coin in the selected box is a gold coin

The desired probability is

$$P\{A_{gg}|B_g\} = \frac{P\{B_g|A_{gg}\} \cdot P\{A_{gg}\}}{P\{B_g\}}$$
$$P\{B_g\} = \tfrac{1}{3} \cdot \tfrac{3}{4} = \tfrac{1}{2}$$
$$P\{A_{gg}\} = \tfrac{1}{3}$$
$$P\{B_g|A_{gg}\} = 1$$
$$P\{A_{gg}|B_g\} = \frac{1 \cdot \frac{1}{3}}{\frac{1}{2}} = \tfrac{2}{3}$$

Example 2-16. In a certain group of engineers, 60 per cent have insufficient background of information theory, 50 per cent have inadequate knowledge of probability, and 80 per cent are in either one or both of the two categories What is the percentage of people who know probability among those who have a

sufficient background of information theory?

Solution. Let

A be those having insufficient background of information theory

B be those having inadequate knowledge of probability

Then

$$\begin{aligned} P\{A\} &= 0.60 & P\{A'\} &= 0.40 \\ P\{B\} &= 0.50 & P\{B'\} &= 0.50 \\ P\{A+B\} &= 0.80 & P\{A+B\}' &= P\{A'B'\} = 0.20 \end{aligned}$$

It is required to find

$$P\{B'|A'\} = \frac{P\{A'B'\}}{P\{A'\}} = \frac{0.20}{0.40} = 50 \text{ per cent}$$

2-11. Theorem of Multiplication.

The multiplication rule for the case of two events A and B can be obtained through the definition of the conditional probability.

$$\begin{aligned} P\{AB\} &= P\{A\}P\{B|A\} \\ P\{AB\} &= P\{B\}P\{A|B\} \end{aligned}$$

(2-83)

This rule can be extended to the case of more than two events. For instance, for three events A, B, and C, one writes

$$\begin{aligned} P\{ABC\} &= P\{AB\}P\{C|AB\} \\ &= P\{A\}P\{B|A\}P\{C|AB\} \end{aligned}$$

(2-84)

More generally,

$$P\{A_1,A_2,\ldots,A_n\} = P\{A_1\}P\{A_2|A_1\}P\{A_3|A_1A_2\}\cdots P\{A_n|A_1,A_2,\ldots,A_{n-1}\}$$

(2-85)

When a finite number or a countably infinite number of events $A_1, A_2, \ldots, A_n$ are mutually independent,[17] we have

$$P\{A_1,A_2,\ldots,A_n\} = P\{A_1\}P\{A_2\}\cdots P\{A_n\}$$

(2-86)

Example 2-17. In a small library there are 1,000 books, among which 500 are scientific. Among the scientific books are 100 which are devoted to engineering subjects. Three books are chosen at random, the chosen book being replaced each time. What is the probability of getting

(*a*)All three scientific books

(*b*)Three scientific books among which only one is an engineering book

(*c*)At least one of the three an engineering book

Solution. Let S and E stand for the event of selecting a scientific and an engineering book, respectively. The events of interest discussed in the problem are

(a) $S_1S_2S_3$

(b) $(S_1E)(S_2E')(S_3E') + (S_1E')(S_2E)(S_3E') + (S_1E')(S_2E')(S_3E)$

(c) $U - E_1'E_2'E_3'$

(a) $P\{S_1S_2S_3\} = P\{S_1\}P\{S_2\}P\{S_3\} = (\frac{1}{2})^3 = \frac{1}{8}$

(b) $P\{SE\} = P\{E|S\}P\{S\}$

$P\{SE\} = \frac{1}{5} \cdot \frac{1}{2} = \frac{1}{10}$

$P\{SE'\} = P\{E'|S\} \cdot P\{S\}$

$P\{SE'\} = \frac{9}{10} \cdot \frac{1}{2} = \frac{9}{10}$

$3P\{(S_1E)(S_2E')(S_3E')\} = 3 \cdot \frac{1}{10} \cdot \frac{9}{10} \cdot \frac{9}{10} = 0.048$

(c) $P\{U - E_1'E_2'E_3'\} = 1 - P\{E_1'E_2'E_3'\}$

$P\{U - E_1'E_2'E_3'\} = 1 - (\frac{9}{10})^3 = 0.271$

Example 2-18. Four persons write their names on individual slips of paper and deposit the slips in a common box. Each of the four draws at random a slip from the box. Determine the probability of each person retrieving his own name slip.

Solution. Let E_k be the event that the *kth* person retrieves his own name slip. The event of interest is $E_1E_2E_3E_4$. Equation (2-85) yields

$$P\{E_1E_2E_3E_4\} = P\{E_1|E_2E_3E_4\} \cdot P\{E_2|E_3E_4\} \cdot P\{E_3|E_4\} \cdot P\{E_4\}$$
$$P\{E_1E_2E_3E_4\} = 1 \cdot \frac{1}{2} \cdot \frac{1}{3} \cdot \frac{1}{4} = \frac{1}{24}$$

Example 2-19. The probability of the closing of each relay of the circuit of Fig. E2-19 is a given . Assuming that all relays act independently, what is the probability of a current existing between terminals *A* and *B*?

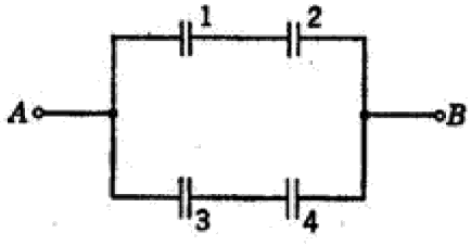


FIG. E2-19

Solution. Let the event of closing each relay 1, 2, 3, and 4 be E_1, E_2, E_3, and E_4, respectively. The four events are independent but not necessarily mutually exclusive. The event of interest is

$$E = E_1E_2 + E_3E_4$$
$$P\{E\} = P\{E_1E_2 + E_3E_4\} = P\{E_1E_2\} + P\{E_3E_4\} - P\{E_1E_2E_3E_4\}$$
$$P\{E\} = P\{E_1\}P\{E_2\} + P\{E_3\}P\{E_4\} - P\{E_1\}P\{E_2\}P\{E_3\}P\{E_4\}$$
$$P\{E\} = 2\alpha^2 - \alpha^4$$

Note that

$$P\{0\} = 0$$
$$P\{1\} = 1$$
$$0 \leq P\{E\} \leq 1 \quad \text{for } 0 \leq \alpha \leq 1$$

2-12. Bayes's Theorem.

In many problems we wish to concentrate on two mutually exclusive and exhaustive events of the sample space, that is, two events A_1 and A $_2$ such that

$$A_1A_2 = \emptyset$$
$$A_1 + A_2 = U$$

(2-87)

The assumption is that each of these events has a subevent of special interest to us. If the subevents are indicated by EA_1 and EA_2, then the event of interest $E = EA_1 + EA_2$ can occur only when A_1 or A_2 occurs. The conditional probabilities $P\{E|A_1\}$ and $P\{E|A_2\}$ are assumed to be known; we are also given the information that E has occurred. The problem is to determine how likely it is that E has occurred because of the occurrence of either of the two events A_1 and A_2. In mathematical notation, given

$$\begin{aligned} P\{A_1\} &= \omega_1 & P\{A_2\} &= \omega_2 \\ A_1 + A_2 &= U & A_1 A_2 &= \emptyset \\ P\{E|A_1\} &= p_1 & P\{E|A_2\} &= p_2 \end{aligned}$$

(2-88)

find $P\{A_1|E\}$ and $P\{A_2|E\}$.

The computation can be done in a direct way by applying the rule of addition and multiplication. Note that

$$E = A_1E \cup A_2E$$

(2-89)

As A_1E and A_2E are mutually exclusive events, we may write

$$P\{E\} = P\{A_1E\} + P\{A_2E\}$$

These probabilities can be calculated as follows:

$$\begin{aligned} P\{A_1E\} &= P\{A_1\}P\{E|A_1\} \\ P\{A_2E\} &= P\{A_2\}P\{E|A_2\} \end{aligned}$$

(2-90)

Therefore,

$$P\{E\} = P\{A_1\}P\{E|A_1\} + P\{A_2\}P\{E|A_2\}$$

$$P\{A_1|E\} = \frac{P\{A_1E\}}{P\{E\}} = \frac{P\{A_1\}P\{E|A_1\}}{P\{A_1\}P\{E|A_1\} + P\{A_2\}P\{E|A_2\}}$$

$$P\{A_2|E\} = \frac{P\{A_2E\}}{P\{E\}} = \frac{P\{A_2\}P\{E|A_2\}}{P\{A_1\}P\{E|A_1\} + P\{A_2\}P\{E|A_2\}}$$

(2-91)

Finally one finds

$$P\{A_1|E\} = \frac{\omega_1 p_1}{\omega_1 p_1 + \omega_2 p_2}$$

$$P\{A_2|E\} = \frac{\omega_2 p_2}{\omega_1 p_1 + \omega_2 p_2}$$

(2-92)

The probabilities expressed in Eqs. (2-92) are called the a posteriori probabilities of A_1 and A_2, given E. The probabilities $_1p_1$ and $_2p_2$ are termed the a priori probabilities of E, given A_1 and A_2. Equations (2-92) provide a means for calculating the a posteriori probabilities from the a priori probabilities. Equations (2-92) are known as Bayes's rule. It is of interest to note that Bayes's rule applies to a partitioned sample space, as shown in Fig. 2-19. The events A_1 and A_2 may each consist of sets containing a number of subevents. Electrical engineers may note that Bayes's rule is somewhat similar to Thévenin's theorem in network theory. Thévenin's theorem permits a partitioning of the network into two parts and a study of the system with respect to one pair of terminals of the partitioned boundary.

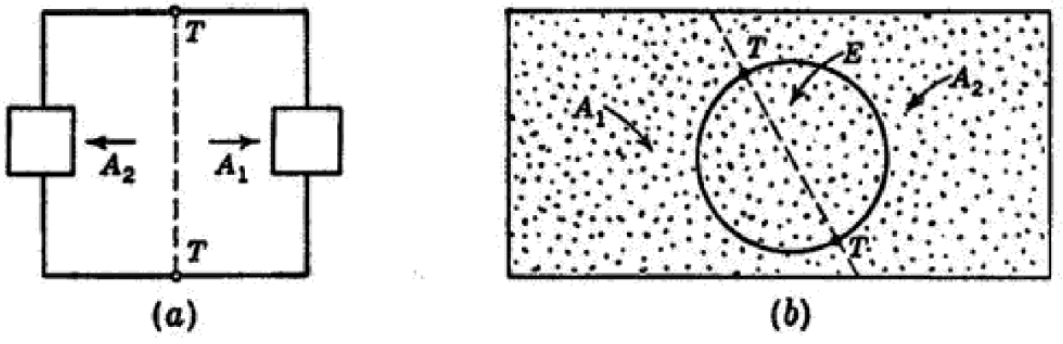


FIG. 2-19. (a) Thévenin's partitioning. A_1, a part of the network; A_2, the remainder network. (b) Bayes's partitioning.

Bayes's theorem, like Thévenin's theorem, can be extended to a partitioning of the sample space into mutually exclusive and exhaustive parts. Suppose that an event E can occur as a result of the occurrence of several mutually exclusive and exhaustive events $A_1, A_2, \ldots, A_n$. Let the corresponding conditional probabilities be given as

$$P\{E|A_k\} = p_k \qquad k = 1, \ldots, n$$

(2-93)

and let

$$P\{A_k\} = \quad _k$$

Then, by the law of addition, we have

$$E = A_1E + A_2E + \cdots + A_nE$$

(2-94)

$$P\{E\} = P\{A_1E + A_2E + \cdots + A_nE\}$$

(2-95)

$$P\{E\} = \omega_1 p_1 + \omega_2 p_2 + \cdots + \omega_n p_n$$

(2-96)

The question is to find the a posteriori probability of the occurrence of event A_k, given the occurrence of E.

$$P\{A_k|E\} = \frac{P\{A_k\}P\{E|A_k\}}{P\{E\}} = \frac{P\{A_k\}P\{E|A_k\}}{\sum_{j=1}^{n} P\{E|A_j\}P\{A_j\}}$$

(2-97)

or, equivalently,

$$P\{A_k|E\} = \frac{\omega_k p_k}{\omega_1 p_1 + \omega_2 p_2 + \cdots + \omega_n p_n}$$

(2-98)

This equation comprises what is known as Bayes's theorem.

Example 2-20. Let U_1, U_2 U_3, be three urns with two red and one black, three red and two black, and one red and one black balls, respectively. One of the three urns is chosen at random and a ball is drawn from it. The color of the ball is found to be black. What is the probability that it has been chosen from U_3?

Solution. This is an example of a situation where Bayes's theorem can be applied. Let E be the event that a black ball has been drawn; A_i is the event that the ith urn has been chosen, $i = 1, 2, 3$.

Then

Also,

$$P\{E|A_1\} = \tfrac{1}{3} \qquad P\{E|A_2\} = \tfrac{2}{5} \qquad P\{E|A_3\} = \tfrac{1}{2}$$

$$P\{A_3|E\} = P\{\text{choosing urn } U_3|\text{black ball drawn}\}$$

$$= \frac{P\{A_3\}P\{E|A_3\}}{\sum_{i=1}^{3} P\{E|A_i\}P\{A_i\}}$$

$$= \frac{\tfrac{1}{3} \cdot \tfrac{1}{2}}{\tfrac{1}{3}(\tfrac{1}{3} + \tfrac{2}{5} + \tfrac{1}{2})} = \frac{15}{37}$$

Example 2-21. Three urns are given:

Urn 1 contains two white, three black, and four red balls.
Urn 2 contains three white, two black, and two red balls.
Urn 3 contains four white, one black, and one red ball.

One urn is chosen at random, and two balls are drawn from that urn. If the two balls happen to be white and red, what is the probability that they were drawn from urn 3?

Solution.

Let A_i = event of choosing urn i, $i = 1, 2, 3$
RW = event of choosing a red and a white ball

We *want* $P\{A_3|RW\}$.

Using Bayes's rule,

But

$$P\{A_3|RW\} = \frac{P\{A_3\}P\{RW|A_3\}}{P\{A_1\}P\{RW|A_1\} + P\{A_2\}P\{RW|A_2\} + P\{A_3\}P\{RW|A_3\}}$$

$$P\{A_1\} = P\{A_2\} = P\{A_3\} = \tfrac{1}{3}$$

$$P\{RW|A_1\} = \frac{8}{\binom{9}{2}} = \frac{8}{36}$$

$$P\{RW|A_2\} = \frac{6}{\binom{7}{2}} = \frac{6}{21}$$

$$P\{RW|A_3\} = \frac{4}{\binom{6}{2}} = \frac{4}{15}$$

Therefore,

$$P\{A_3|RW\} = \frac{\tfrac{1}{3} \cdot \tfrac{4}{15}}{\tfrac{1}{3}(\tfrac{3}{9} + \tfrac{6}{21} + \tfrac{4}{15})} = \frac{21}{61}$$

Bayes's[18] theorem comprises one of the most used, and occasionally misused, concepts of probability theory. In many problems an event may occur as an "effect" of several "causes." From a number of observations on the occurrence of the effect, one can make an estimate on the occurrence of the cause leading to that effect. This rule is frequently applied to communication problems, particularly in the detection of signals from an additive mixture of signals and noise. When the detecting instrument indicates a signal, we have to make a decision whether the received signal is a true one or a false alarm due to undesired signals (noise) in the system. Such decisions are generally made possible by an application of Bayes's rule which is also called the rule of inverse probability. The decision criterion may be made more effective by introducing some kind of weighting coefficients called *loss matrix* and minimizing the overall "loss."

2-13. Combinatorial Problems in Probability.

In many problems involving choice and probability, the number of possible ways of arranging a given number of objects on a line is of interest. For example, if three persons *A*, *B*, and *C* are standing in a line, the probability that *A* remains next to *B* can be calculated as follows: There are six different arrangements possible:

ABC ACB BAC BCA CAB CBA

Of these arrangements, there are four desirable ones. Thus, if the concept of equiprobable measures is assumed, the probability in question is $\frac{2}{3}$.

Combinatorial problems have a limited use in our subsequent studies. For this reason, we shall give only a review of the most pertinent definitions in this section. The reader interested in combinatorial problems will find a considerable amount of information in Feller (Chaps. 2 to 4).

Permutation: A permutation of the elements of a finite set is a one-to-one correspondence between elements of that set (such a correspondence is also called a mapping of the set onto itself). For example, if a set contains only four objects *A*, *B*, *C*, and *D*, we may write two equivalent sets

A, B, C, D and *B, C, A, D*

The ordered sets; $\begin{bmatrix} A & B & C & D \\ 1 & 2 & 3 & 4 \end{bmatrix}$ and $\begin{bmatrix} B & C & A & D \\ 1 & 2 & 3 & 4 \end{bmatrix}$ are two permutations of the elements of the original set, since

$$A \rightarrow B$$
$$B \rightarrow C$$
$$C \rightarrow A$$
$$D \rightarrow D$$

(2-99)

The following definition is of considerable assistance in dealing with combinatorial problems.

Factorial: The factorial function for a positive integer n is defined as

$$n! = n(n-1)(n-2) \cdots 4 \cdot 3 \cdot 2 \cdot 1$$

(2-100)

with the additional convention

$$0! = 1$$

(2-101)

The number of different permutations of a set with n distinct elements is

$$P_n = n(n-1)(n-2) \cdots 4 \cdot 3 \cdot 2 \cdot 1 = n!$$

(2-102)

Combination: The number of different permutations of r objects selected from n objects is

$$P_r^n = n(n-1)(n-2)\cdots(n-r+1) = \frac{n!}{(n-r)!}$$

(2-103)

Every permutation of elements of a set contains the same elements but in different order. When two sets of objects are in one-to-one correspondence so that some of the elements of one do not appear in the other they are called different *combinations.* For example, if we combine the members of the set {*A,B,C,D*} two by two, *AB, AC, DB* are different combinations but *AB* and *BA* are not.

The number of different combinations of n objects taken r at a time is

$$C_r^n = \frac{P_r^n}{r!} = \frac{n!}{r!(n-r)!}$$

(2-104)

When confusion will not result, one may use the notation $\binom{n}{r}$ for C_r^n.

Note that

$$\binom{n}{n-r} = \binom{n}{r}$$

(2-105)

$$\binom{n}{1} = n$$

(2-106)

$$\binom{n+1}{r} = \binom{n}{r-1} + \binom{n}{r}$$

(2-107)

$$\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n} = \sum_{r=0}^{n} \binom{n}{r} = 2^n$$

(2-108)

The following theorem is often used in combinatorial problems. Let a set contain k mutually exclusive subsets of objects: with

$$\begin{matrix} \{A_1, A_2, \ldots, A_k\} \\ A_i = \{a_{i1}, a_{i2}, \ldots, a_{in_i}\} \quad i = 1, 2, \ldots, k \end{matrix}$$

n_i being the number of elements in the set A_i. The number of permutations of the total number of elements n is

$$\frac{n!}{n_1! n_2! \cdots n_k!}$$

(2-109)

In fact, one has to divide the number of permutations of n objects by $n_i!$ (for $i = 1, 2, \ldots, k$) since the permutations of the identical objects of the A_i set cannot be distinguished from each other. For example, the number of color permutations of three black and two white balls is

$$\frac{5!}{3!2!} = \frac{5 \times 4}{2!} = 10$$

Binomial Expansion: Let n be a positive integer; then

$$(a+b)^n = a^n + \binom{n}{1} a^{n-1}b + \binom{n}{2} a^{n-2}b^2 + \cdots + \binom{n}{r} a^{n-r}b^r + \cdots + b^n$$

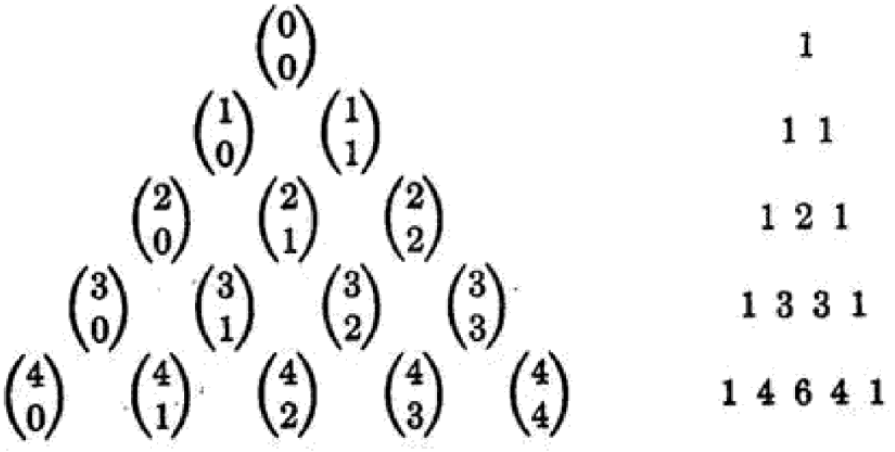
(2-110)

or

$$(a+b)^n = a^n + na^{n-1}b + \frac{n(n-1)}{2!} a^{n-2}b^2 + \frac{n(n-1)(n-2)}{3!} a^{n-3}b^3 + \cdots + b^n$$

(2-111)

A useful of a binomial coefficent is given in a table which is called *Pascal triangle:*



(2-112)

In the following a number of simple examples dealing with permutations and combinations are presented. In these examples, the primary assumption is that the probability is given by the frequency of the event under consideration; that is, the concept of equiprobable measure prevails. Hence, such problems are reduced to a study of the ratio of the favorable cases to all possible cases. In this respect the formula of combinatorial analysis will be used.

Example 2-22. What is the probability of a person having four aces in a bridge hand?

Solution. The number of all possible different hands equals the combination of 13 from 52 cards. For the number of favorable cases one may think of first removing the four aces from the deck and then dealing all possible combinations of hands 9 by 9. The addition of the four aces to each one of these latter hands gives a favorable case.

$$\binom{48}{9} : \binom{52}{13} = \frac{10 \cdot 11 \cdot 12 \cdot 13}{49 \cdot 50 \cdot 51 \cdot 52} = \frac{11}{4{,}165}$$

Example 2-23. Two cards are drawn from a regular deck of cards. What is the probability that neither is a heart?

Solution. Let A and B be the events that the first and the second card are hearts, respectively; then we wish to know $P\{A \quad B \quad\}$.

Therefore

$$P\{A'\} = 1 - P\{A\} = 1 - \tfrac{1}{4} = \tfrac{3}{4}$$
$$P\{B'|A'\} = \frac{P\{A'B'\}}{P\{A'\}}$$
$$P\{B'|A'\} = \tfrac{38}{51}$$
$$P\{A'B'\} = \tfrac{38}{51} \cdot \tfrac{3}{4} = \tfrac{19}{34}$$

If we wish to apply combinatorial principles, we may say that the number of all possible cases of selecting two cards is $\binom{52}{2}$. The number of favorable cases is $\binom{39}{2}$.

Therefore the probability in question is

$$\binom{39}{2} : \binom{52}{2} = \frac{39!}{2!37!} \frac{2!50!}{52!} = \frac{39 \cdot 38}{51 \cdot 52} = \frac{19}{34}$$

2-14. Trees and State Diagrams.

The material of this section is intended to offer a graphical interpretation for certain simple problems of probability which arise in dealing with repeated trials of an experiment.

For example, suppose that a biased coin is tossed once; the outcome may be denoted by H and T and shown by the diagram of Fig. 2-20. Similarly, if the coin is tossed twice, the second set of outcomes may be shown in the same treelike diagram. If the probability of getting a head is denoted by p, then the probability of getting, say, HT can be directly computed from the weighted length of the associated tree path, that is,

$$p(1 - p)$$

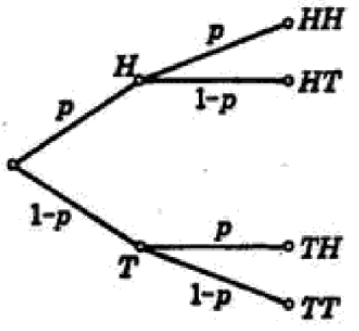


FIG. 2-20. A simple tree diagram.

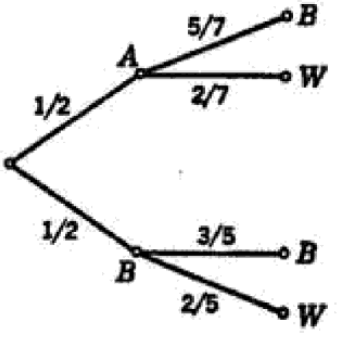


FIG. 2-21. An example of a probability tree.

If it is desired to obtain the probability of getting a head and a tail irrespective of their order, then the answer to the problem is given by summing up the two weighted tree paths.

$$p(1 - p) + (1 - p)p = 2p(1 - p)$$

This simple graphical procedure can be used profitably in certain types of problems. The following are examples of such problems.

Example 2-24. The urn *A* contains five black and two white balls. The urn *B*

contains three black and two white balls. If one urn is selected at random, what is the probability of drawing a white ball from that urn?

Solution. From the tree diagram of Fig. 2-21 one can see that the probability of the event of interest is the sum of the following measures:

$$\tfrac{1}{2} \cdot \tfrac{3}{7} + \tfrac{1}{2} \cdot \tfrac{3}{5} = \tfrac{13}{35}$$

Example 2-25. Find the probability that at least three heads are obtained in a sequence of four throws of an honest coin.

Solution. From a tree diagram or from the binomial expansion one obtains

$$\binom{4}{4} \cdot (\tfrac{1}{2})^4 + \binom{4}{3} \cdot (\tfrac{1}{2})^3(\tfrac{1}{2})^1 = \tfrac{1}{16} + \tfrac{4}{16} = \tfrac{5}{16}$$

If a coin is tossed n times, we note that the probability of getting, say, exactly r heads ($r < n$) is the sum of the tree measures of all tree paths leading to r heads and $n - r$ tails. Since there are $\binom{n}{r}$ such states, it is found that the desired probability is

$$\binom{n}{r} p^r \cdot (1 - p)^{n-r}$$

(2-113)

The tree diagram can easily be drawn for experiments with a finite number of outcomes. In the problems discussed thus far in this section, it is tacitly assumed that the outcomes of each experiment remain independent of the previous experiments. In engineering terminology such experiments are said to lack memory. For these experiments the probability of any outcome is always the same. That is, an outcome of the nth trial has exactly the same probability of

occurrence as in the kth trial ($k \neq n$). n). This type of experiment leads to the concept of so-called *independent stochastic processes.* In certain types of problems an outcome may be influenced by the past history or "memory" of the experiment. *Such* experiments are termed *dependent stochastic processes.* Among the latter type, perhaps the simplest ones are those experiments in which the probability of an outcome of a trial depends on the outcome of the immediately preceding trial. These are called *Markov processes.*

Let an experiment have a finite number of n possible outcomes, $a_1, a_2, \ldots,$ and a_n, called *states.* We assume the process to be of the finite Markov type and initially in the state k. For a Markov process, we specify a table of probabilities associated with transitions from any state to any other state. This is called a *probability transition matrix.*

$$\begin{array}{c} \\ a_1 \\ a_2 \\ a_3 \\ \cdots \\ a_n \end{array} \begin{array}{c} \begin{array}{cccccc} a_1 & a_2 & a_3 & \cdots & a_n \end{array} \\ \begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots & p_{1n} \\ p_{21} & p_{22} & p_{23} & \cdots & p_{2n} \\ p_{31} & p_{32} & p_{33} & \cdots & p_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ p_{n1} & p_{n2} & p_{n3} & \cdots & p_{nn} \end{bmatrix} \end{array}.$$

(2-114)

$p_{jk} = p\{a_k|a_j\}$ denotes the probability that the next outcome of the experiment will be the state k, given that the immediately preceding experiment led to the state j. Note that in a transition probability matrix the sum of all elements of each row must equal unity.

One of the most common problems associated with the Markov process is, given that it started with state j, to find the probability of reaching the state k after a specified number of steps r, that is, $p\{a_k|a_j\}^{(r)} = p_{jk}^{(r)}$. This question has a rather simple answer, namely, (1) draw the tree diagram, (2) select all tree paths connecting the node representing the state j to that of the state k in r steps, and (3) add the corresponding tree measures. This procedure is exemplified in the

tree diagram of Fig. 2-22 for $r = 1$ and $r = 2$. When $r = 1$, the answer is obvious:

$$P\{a_k|a_j\}^{(1)} = P\{a_j\}P\{a_k|a_j\}$$

(2-115)

For $r = 2$ one has to add the probabilities of reaching state a_k from the state a_j in all possible ways, that is, the sum of the measures associated with all three paths connecting a_j to a_k in two steps.

$$\begin{aligned} P\{a_k|a_j\}^{(2)} &= P\{a_j\}[P\{a_1|a_j\}P\{a_k|a_1\} + P\{a_2|a_j\}P\{a_k|a_2\} \\ &\qquad + \cdots + P\{a_n|a_j\}P\{a_k|a_n\}] \\ &= P\{a_j\}\sum_{i=1}^{i=n} P\{a_i|a_j\}P\{a_k|a_i\} = P\{a_j\}\sum_{i=1}^{n} P_{ji}\cdot P_{ik} \end{aligned}$$

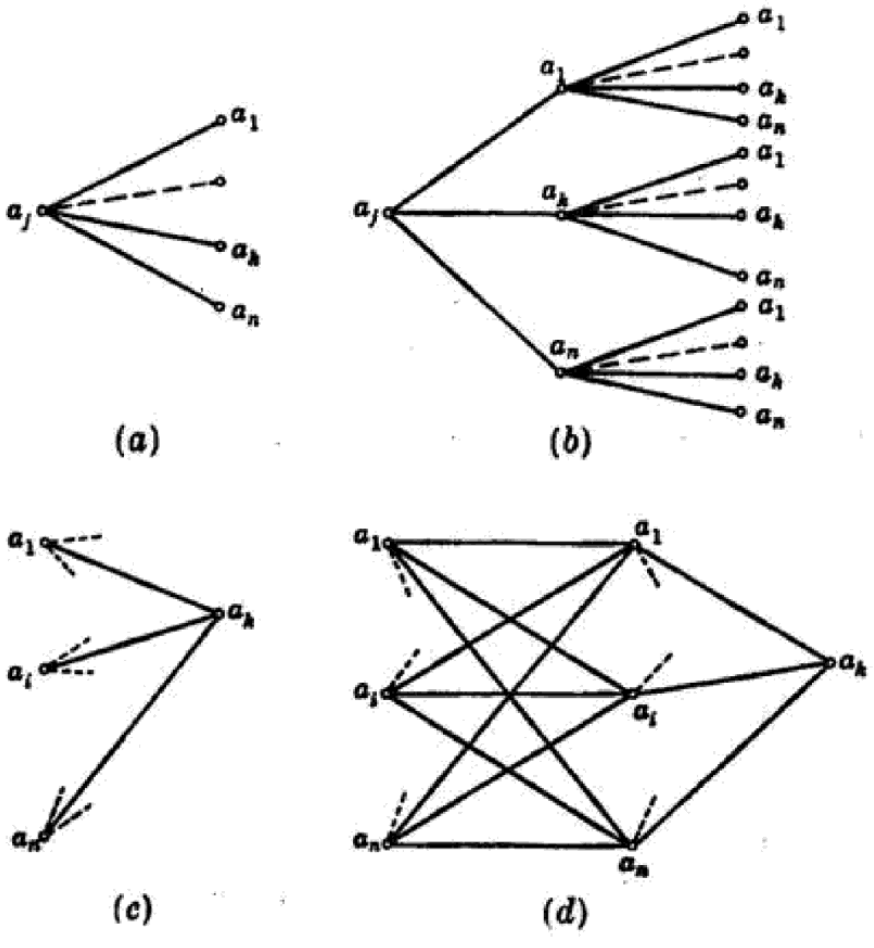


FIG. 2-22. Trees for a finite chain. (a) r = 1. (b) r = 2. (c) r = 1. (d) r = 2.

For $r = 3$,

$$P\{a_k|a_j\}^{(3)} = P\{a_j\}\sum_{\rho=1}^{\rho=n}\sum_{i=1}^{i=n} P\{a_i|a_j\}P\{a_\rho|a_i\}P\{a_k|a_\rho\}$$

By defining the initial probability of different states as a diagonal matrix,

$$[P_D^{(0)}] = \begin{bmatrix} P\{a_1\} & 0 & \cdots & 0 \\ 0 & P\{a_2\} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & P\{a_n\} \end{bmatrix}$$

we can sum up the above development in concise matrix notation. That is, for any states j and k we have

$$[P\{a_j|a_k\}^{(1)}] = [P_D^{(0)}][P]$$

Similarly,

$$[P\{a_j|a_k\}^{(2)}] = [P_D^{(0)}][P][P] = [P_D^{(0)}][P]^2$$

For the general case,

$$[P\{a_j|a_k\}^{(r)}] = [P_D^{(0)}][P]^r$$

(2-116)

This relation determines the probability $P\{a_j|a_k\}^{(r)}$ for any values of j, k, and r.

Consider next the probability of reaching the state a_k in r steps, given that the initial state could have been any a_i, $i = 1, 2, \ldots, n$, that is, the probability of getting to a_k (in r steps) when any of the n states could have been the initial state. Let this probability be symbolized by $P\{a_k|\ \}^{(r)}$. Figure 2-22c and d illustrates the case for $r = 1$ and $r = 2$, respectively.

For $r = 1$,

$$P\{a_k|\quad\}^{(1)} = \sum_{i=1}^{n} P\{a_i\}P\{a_k|a_i\}$$

For $r = 2$,

$$P\{a_k|\quad\}^{(2)} = \sum_{i=1}^{n}\sum_{\varrho=1}^{\varrho=n} P\{a_i\}P\{a_\varrho|a_i\}P\{a_k|a_\varrho\}$$

For $r = 3$,

$$P\{a_k|\quad\}^{(3)} = \sum_{i=1}^{n}\sum_{\varrho=1}^{n}\sum_{\lambda=1}^{n} P\{a_i\}P\{a_\varrho|a_i\}P\{a_\lambda|a_\varrho\}P\{a_k|a_\lambda\}$$

The matrix formulation follows immediately. Let $[P^0]$ be a row matrix describing the initial probabilities $[P\{a_1\}, P\{a_2\}, \ldots, P\{a_n\}]$; then

$$[P\{a_k|\quad\}^{(1)}] = [P^{(0)}][P]$$
$$[P\{a_k|\quad\}^{(2)}] = [P^{(0)}][P]^2$$

For the general case,

$$[P\{a_k|\quad\}^{(r)}] = [P^{(0)}][P]^r$$

(2-117)

This relation determines the probability $P\{a_k|\}^{(r)}$ for any values of positive integers k and r. Note that $P\{a_k|\}^{(r)}$ will always be a row matrix since $[P^{(0)}]$ is a row matrix.

Example 2-26. A relay alternates between the open state denoted by 1 and the closed state designated by 0. The transition probability matrix is given as

$$\begin{array}{c} \\ 1 \\ 0 \end{array} \begin{array}{c} \begin{array}{cc} 1 & 0 \end{array} \\ \begin{bmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{2}{3} \end{bmatrix} \end{array}$$

Assuming that the initial probability of the relay being in either state is ½, determine

(*a*)The probability of reaching state 1 via state 0 in one step, that is, $p_{01}^{(1)}$.

(*b*)$p_{00}^{(1)}$.

(*c*)$p_{01}^{(2)}$,

(*d*)$p_{11}^{(2)}$.

(*e*)$p\{1|\ \}^{(2)}$, the probability of reaching state 1 in two steps.

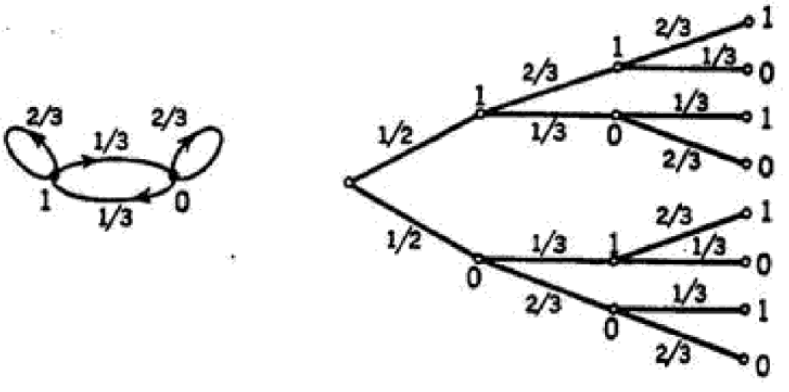


FIG. E2-26

Solution. The state diagram and the tree diagram are drawn in Fig. E2-26. According to the tree diagram,

$$
\begin{aligned}
&(a) && p_{01}^{(1)} = \tfrac{1}{2} \cdot \tfrac{1}{3} = \tfrac{1}{6} \\
&(b) && p_{00}^{(1)} = \tfrac{1}{2} \cdot \tfrac{2}{3} = \tfrac{1}{3} \\
&(c) && p_{01}^{(2)} = \tfrac{1}{2}(\tfrac{1}{3} \cdot \tfrac{2}{3} + \tfrac{2}{3} \cdot \tfrac{1}{3}) = \tfrac{2}{9} \\
&(d) && p_{11}^{(2)} = \tfrac{1}{2}(\tfrac{2}{3} \cdot \tfrac{2}{3} + \tfrac{1}{3} \cdot \tfrac{1}{3}) = \tfrac{5}{18} \\
&(e) && p\{1|\ \ \}^{(2)} = p_{11}^{(2)} + p_{01}^{(2)} = \tfrac{5}{18} + \tfrac{2}{9} = \tfrac{1}{2}
\end{aligned}
$$

An alternative solution for part (*e*) is given by the matrix relation of Eq. (2-117).

$$
[\tfrac{1}{2} \quad \tfrac{1}{2}] \begin{bmatrix} \tfrac{2}{3} & \tfrac{1}{3} \\ \tfrac{1}{3} & \tfrac{2}{3} \end{bmatrix}^2 = [\tfrac{1}{2} \quad \tfrac{1}{2}] \begin{bmatrix} \tfrac{5}{9} & \tfrac{4}{9} \\ \tfrac{4}{9} & \tfrac{5}{9} \end{bmatrix} = [\tfrac{1}{2} \quad \tfrac{1}{2}]
$$

Therefore,

$$
p\{1|\ \ \}^{(2)} = p\{0|\ \ \}^{(2)} = \tfrac{1}{2}
$$

Finally we may answer the same question by using the materials of Sees. 2-9 (Theorem of Addition) and 2-10 (Conditional Probability).

$$
\begin{aligned}
&(a) && p\{01\} = p\{0\}p\{1|0\} = \tfrac{1}{2} \cdot \tfrac{1}{3} = \tfrac{1}{6} \\
&(b) && p\{00\} = p\{0\}p\{0|0\} = \tfrac{1}{2} \cdot \tfrac{2}{3} = \tfrac{1}{3} \\
&(c) && p\{001\} + p\{011\} = p\{00\}p\{1|0\} + p\{01\}p\{1|1\} \\
& && \qquad = \tfrac{1}{3} \cdot \tfrac{1}{3} + \tfrac{1}{6} \cdot \tfrac{2}{3} = \tfrac{2}{9} \\
&(d) && p\{111\} + p\{101\} = p\{11\}p\{1|1\} + p\{10\}p\{1|0\} \\
& && \qquad = \tfrac{1}{2} \cdot \tfrac{2}{3} \cdot \tfrac{2}{3} + \tfrac{1}{2} \cdot \tfrac{1}{3} \cdot \tfrac{1}{3} = \tfrac{5}{18} \\
&(e) && p\{1|\ \ \}^{(2)} = \tfrac{2}{9} + \tfrac{5}{18} = \tfrac{1}{2}
\end{aligned}
$$

Example 2-27. A communication source having a three-letter alphabet transmits sequences of messages. The transition probability matrix is given below:

$$
\begin{array}{c} \\ A \\ B \\ C \end{array}
\begin{array}{c}
\begin{array}{ccc} A & B & C \end{array} \\
\begin{bmatrix} 0 & \tfrac{1}{3} & \tfrac{2}{3} \\ \tfrac{1}{3} & \tfrac{1}{3} & \tfrac{1}{3} \\ \tfrac{5}{9} & 0 & \tfrac{4}{9} \end{bmatrix}
\end{array}
$$

For the beginning of each message, letters A, B, C occur with probabilities $\frac{3}{18}$, $\frac{8}{18}$, and $\frac{7}{18}$, respectively.

(a) Determine the probability of getting a message commencing with

$$AB,\quad BB,\quad CA,$$
$$ABA,\quad BBC,\quad CAC$$

(b) Find a set of initial probabilities which will produce a so-called "steady state," i.e., the probability that the letter transmitted at the nth state does not depend on n.

Solution

(a) The probabilities in question are, respectively,

$$\frac{3}{18}\cdot\frac{1}{3}=\frac{3}{54} \qquad \frac{8}{18}\cdot\frac{1}{3}=\frac{8}{54} \qquad \frac{7}{18}\cdot\frac{5}{9}=\frac{35}{162}$$
$$\frac{3}{54}\cdot\frac{1}{3}=\frac{3}{162} \qquad \frac{8}{54}\cdot\frac{1}{3}=\frac{8}{162} \qquad \frac{35}{162}\cdot\frac{2}{3}=\frac{35}{243}$$

(b) The desired initial probability matrix $[P^{(0)}] = [\ ,\ ,\ ]$ must satisfy the condition

$$[P^{(0)}][P]^{(n)} = [P^{(0)}][P]^{(n-1)} \quad n \text{ a positive integer}$$

In particular,

$$[P^{(0)}][P] = [P^{(0)}]$$

Therefore

$$\alpha = \beta\tfrac{1}{3} + \gamma\tfrac{5}{9}$$
$$\beta = \alpha\tfrac{1}{3} + \beta\tfrac{1}{3}$$
$$\gamma = \alpha\tfrac{2}{3} + \beta\tfrac{1}{3} + \gamma\tfrac{4}{9}$$

These equations lead to

$$\alpha = \tfrac{1}{3} \qquad \beta = \tfrac{1}{6} \qquad \gamma = \tfrac{1}{2}$$

It can be shown that, if one considers very long messages, the frequency of the occurrence of the letter A will approach $\frac{1}{3}$, etc. For further comments on the Markov chain, see Chap. 11.

2-15. Random Variables.

In the preceding sections the concept of an event and of sample space of an experiment played an important role. The discussion of the present section is aimed at an intuitive introduction of *random variables.*

Most experiments of practical interest have numerical outcomes; that is, the result of the experiment is a number, or a pair of numbers, etc. In other words, the results can be described by using a coordinate space, the coordinate space being in a correspondence with the sample space of the event.

A random variable is a real-valued function defined over the sample space of a random experiment. Restricting the random variable to assume only real values is quite natural, as one is interested in the numerical outcomes of an experiment (even though in various practical applications complex values of random variables are also considered). The word "random" stresses the fundamental fact that we are dealing with experiments governed by laws of chance rather than any deterministic law. The throws of a symmetrical die or coin under hypothetically symmetrical conditions represent random experiments. The salient feature of these experiments is that, even though they exhibit a certain kind of regularity when repeated over a long range of time, it is impossible to predict, with complete certainty, the outcome of any particular trial.

Let Ω be the sample space of a random experiment. Each point of Ω describes a possible outcome of the experiment. This outcome may not be a numerical result in itself but some numerical data can be assigned to it. For instance, if the experiment were the picking at "random" of a card out of a deck of 52, the number of possible outcomes at any particular trial would be 52, depending

upon which one of the cards had been picked. Here, although the outcome does not furnish us with a numerical result, we can represent the possible outcomes by, say, the first 52 integers or by 52 points on a line.

The correspondence between a point of Ω and a point in the coordinate space is designated by a mathematical function. This function is termed a *random variable.* Generally, we shall denote random variables by capital letters such as X and Y, and their specific values by the same letters in lower case. A random variable X assumes different values $x_1, x_2, \ldots, x_n, \ldots$ which are points of the coordinate space. The coordinate space may be a one-dimensional or a multidimensional space. The random variable may take a finite number of n-tuple values or infinitely many. The sample space may be a space with finite or countably infinite points or even a continuous space, that is, with an uncountable number of points. The following practical examples illustrate some possibilities.

Example 2-28. The experiment is throwing an ordinary honest die. The sample space has six events of interest. The associated random variable takes only six possible numerical values, 1, 2, 3, 4, 5, and 6. Each of these real numbers corresponds to a specific event.

Example 2-29. The experiment is throwing three honest dice. The associated random variable takes on 6^3 different numbers of triads as values. The random variable may be conveniently represented by a point in the three-dimensional euclidean space.

2-16. Discrete Probability Functions and Distributions.

Consider Ω the sample space of a random experiment. If the outcomes of this experiment can be put into one-to-one correspondence with the positive integers, the sample space will contain a countable number of points. Such a sample space is said to be a *discrete sample space.* In a discrete sample space, when the random variable X assumes values

$$[x_1, x_2, \ldots, x_k, \ldots\}$$

the probability function $f(x)$ is defined as where

$$[p_1,p_2, \ldots ,p_k, \ldots]$$
$$f(x_k) = P\{X = x_k\} = p_k$$

(2-118)

The *probability distribution function* $F(x)$, known also as the *cumulative distribution function* (CDF), is defined as

$$F(x) = \sum_{x_j \leq x} f(x_j)$$

(2-119)

For example, the throw of an honest coin until a head appears is a random experiment. The sample space of this experiment is a discrete space. If X corresponds to the event of the appearance of the first head on the kth throw, then X assumes the following values:

$$[X] = [1,2,3, \ldots ,k, \ldots]$$

(2-120)

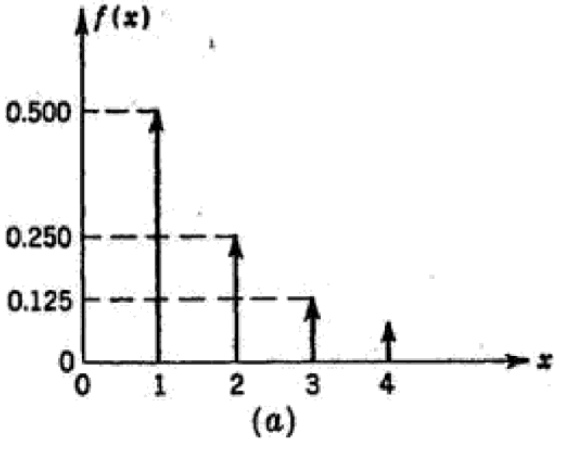
The probability function $f(x)$ and the CDF are

$$f(x) = [2^{-1},2^{-2},2^{-3}, \ldots ,2^{-k}, \ldots]$$
$$F(x) = 2^{-1} + 2^{-2} + \cdots + 2^{-x}$$

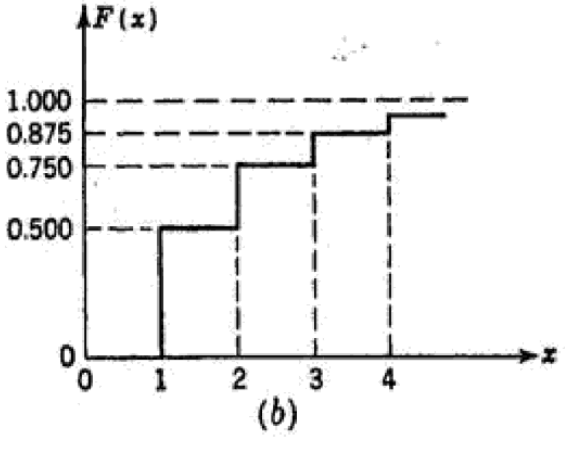
These functions are plotted in Fig. 2-23*a* and *b*, respectively.

The definition of the probability function and CDF can be directly extended to the case of a multivariate random variable. For instance, in Example 2-29 the sample space is a three-dimensional euclidean space with 216 points. The random variable X assumes 216 triad values $X = (X_1,X_2,X_3)$ for any experiment. The corresponding probability function is

$$f(x_1,x_2,x_3) = P\{X_1 = x_1, X_2 = x_2, X_3 = x_3\}$$
$$f(x_1,x_2,x_3) = \tfrac{1}{6} \cdot \tfrac{1}{6} \cdot \tfrac{1}{6} = \tfrac{1}{216}$$



(*a*)



(*b*)

FIG. 2-23. (a) Probability function associated with Eq. (2-121). (b) CDF associated with Eq. (2-121).

Here all permissible outcomes have equal probabilities.

The CDF gives the total probability of the set of points having each coordinate less than or equal to some specified value (x_1,x_2,x_3), that is,

$$F(x_1,x_2,x_3) = \sum f(x_i,x_j,x_k)$$

for

$$x_i \le [x_1] \quad x_j \le [x_2] \quad x_k \le [x_3]$$

where [ ] denotes the greatest integer contained in the letter inside the brackets.

2-17. Bivariate Discrete Distribution.

The case of a random variable assuming pairs of values (x_j,Y_k) is of particular interest. In fact, in most engineering problems the interrelation between two random quantities leads to a bivariate discrete distribution. The joint probability function and the CDF are defined as before:

$$f(x,y) = P\{X = x,\ Y = y\}$$
$$F(x,y) = P\{X \le x,\ Y \le y\}$$

(2-122)

If the joint probability function $f(X,Y)$ is known, say in the form of a matrix, then there are four additional quantities of interest which can be readily computed. These are marginal probability functions and marginal CDF's as defined below:

$$f_1(x_i) = P\{X = x_i, \text{ all permissible } Y\text{'s}\} = \sum_y f(x_i, y)$$
$$f_2(y_j) = P\{Y = y_j, \text{ all permissible } X\text{'s}\} = \sum_x f(x, y_j)$$
$$F_1(x_i) = \sum_{x_k \leq x_i} f_1(x_k)$$
$$F_2(y_j) = \sum_{y_k \leq y_j} f_2(y_k)$$

(2-123)

The indices 1 and 2 in the marginal distributions are simply to indicate that $f_1(x)$ refers to the variable x, that is, the first variable, and $f_2(y)$ to the second variable. Now assume that all pairs of values (x_i, Y_j) are written in a matrix form:

$$[X,Y] = \begin{bmatrix} (x_1,y_1) & (x_1,y_2) & \cdots & (x_1,y_n) \\ (x_2,y_1) & (x_2,y_2) & & (x_2,y_n) \\ \cdots & \cdots & \cdots & \cdots \\ (x_m,y_1) & (x_m y_2) & & (x_m,y_n) \end{bmatrix}$$

(2-124)

The corresponding probabilities can be written in a similar form:

$$[f(x_i,y_j)] = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ p_{m1} & p_{m2} & & p_{mn} \end{bmatrix}$$

(2-125)

The marginal probability $f_1(x_2)$ is the probability of the occurrence of events for which $X = x_2$ without regard to the value of Y. This is readily obtained by adding the terms appearing in the second row of the probability matrix.

$$f_1(x_2) = p_{21} + p_{22} + \cdots + p_{2n}$$

(2-126)

Similarly, the marginal distribution $f_2(y_k)$ can be obtained by adding the terms of the kth column of the joint probability matrix. For example,

$$f_2(y_2) = p_{12} + p_{22} + \cdots + p_{m2}$$

(2-127)

If the random variables X and Y are such that for all values of (x_i,y_j) we have

$$f(x_i,y_j) = f_1(x_i)f_2(y_j)$$

(2-128)

then the variables are said to be statistically independent of each other. For example, the simultaneous throw of two honest coins has the following outcomes:

$$[X,Y] = \begin{bmatrix} H_1H_2 & H_1T_2 \\ T_1H_2 & T_1T_2 \end{bmatrix} \quad \text{and} \quad [f(x,y)] = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}$$

Evidently, these two variables are independent of each other, since for any entry of the probability matrix we have

$$P\{X = H_1\} = p_{11} + p_{12} = \tfrac{1}{2}$$
$$P\{Y = T_2\} = p_{12} + p_{22} = \tfrac{1}{2}$$
$$P\{X = H_1, Y = T_2\} = (p_{11} + p_{12})(p_{12} + p_{22}) = \tfrac{1}{2} \cdot \tfrac{1}{2} = \tfrac{1}{4}$$

Conversely, a check for independence is to determine if Eq. (2-128) holds for all possible outcomes.

The conditional distributions can also be defined and obtained in a straightforward manner. The conditional probability $P\{X = x_i | Y = y_j\}$ is designated as $f(x_i|y_j)$. That is, if the computation of $f(x_i|y_j)$ is desired, then we concentrate on the jth column of the (x,y) matrix.

$$[X, Y = y_j] = \begin{bmatrix} x_1 y_j \\ x_2 y_j \\ \cdots \\ x_i y_j \\ \cdots \\ x_m y_j \end{bmatrix}$$

(2-129)

Next the term $x_i y_j$ is selected and its associated probability is obtained.

$$P\{X = x_i | Y = y_j\} = f(x_i|y_j) = \frac{f(x_i, y_j)}{f_2(y_j)}$$
$$f_2(y_j) \neq 0$$

(2-130)

It is to be noted that $f(x_i|y_j)$ is a permissible conditional probability function as all its terms are nonnegative and

$$\sum_{i=1}^{m} f(x_i|y_j) = \frac{\sum_{i=1}^{m} f(x_i,y_j)}{f_2(y_j)} = \frac{f_2(y_j)}{f_2(y_j)} = 1$$
$$f_2(y_j) \neq 0$$

(2-131)

Similarly, the conditional probability of Y, given $X = x_i$, is found to be

$$f(y_j|x_i) = \frac{f(x_i,y_j)}{f_1(x_i)}$$
$$f_1(x_i) \neq 0$$

(2-132)

Example 2-30. Consider the simultaneous throw of two honest dice X and Y. Find $P\{3 \leq X \leq 5, 2 \leq Y \leq 3\}$ and the marginal probabilities.

Solution. The two-dimensional random variable assumes 36 pairs of values, each with an equal probability of $\frac{1}{36}$.

$$P_{jk} = \tfrac{1}{36} \quad \text{for each point of the sample space}$$
$$F(x,y) = P(X \leq x, Y \leq y)$$

The marginal CDF's are

$$F_1(x) = \sum_{j=1}^{[x]} \sum_{k=1}^{6} p(j,k) = \frac{[x]}{6}$$

$$F_2(y) = \sum_{k=1}^{[y]} \sum_{j=1}^{6} p(j,k) = \frac{[y]}{6}$$

Y \ X	1	2	3	4	5	6	$f_1(x)$
1	1/36	1/36	1/36	1/36	1/36	1/36	1/6
2	1/36	1/36	1/36	1/36	1/36	1/36	1/6
3	1/36	1/36	1/36	1/36	1/36	1/36	1/6
4	1/36	1/36	1/36	1/36	1/36	1/36	1/6
5	1/36	1/36	1/36	1/36	1/36	1/36	1/6
6	1/36	1/36	1/36	1/36	1/36	1/36	1/6
$f_2(y)$	1/6	1/6	1/6	1/6	1/6	1/6	

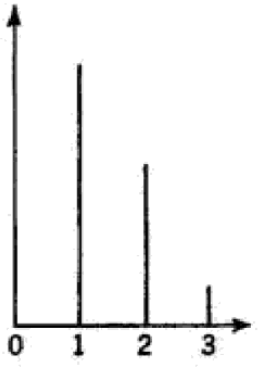
The probability of having $3 \le X \le 5$ and $2 \le Y \le 3$ is $\frac{6}{36} = \frac{1}{6}$. The marginal probabilities are $P\{3 \le X \le 5\} = \frac{1}{2}$ and $P\{2 \le Y \le 3\} = \frac{1}{3}$. Note that the two variables are independent, since, for all entries of the probability matrix, $\frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6}$.

2-18. Binomial Distribution.

Consider a random experiment with only two possible outcomes, E_1 and E_2. Let the probability of the occurrence of E_1 and E_2 be p and $q = 1 - p$, respectively. If the experiment is, repeated n times and the successive trials are independent of each other, the probability of obtaining E_1 and E_2 r and $n - r$ times, respectively, is

$$\binom{n}{r} p^r q^{n-r}$$

This can be proved as follows: The probability of any sequence having r events E_1 and $n - r$ events E_2 is $p^r q^{n-r}$, as the successive trials are assumed to be independent of each other. Moreover, the number of such sequences is equal to the number of combinations of n objects r at a time. Hence the formula of Eq. (2-133) follows by the addition rule of probabilities.



r	$\binom{n}{r}$	p^r	q^{n-r}	$\binom{n}{r} p^r q^{n-r}$
0	1	1.000	0.216	0.216
1	3	0.400	0.360	0.432
2	3	0.160	0.600	0.288
3	1	0.064	1.000	0.064

FIG. 2-24. Example of a binomial probability function.

Let us now define a random variable X which takes the values r if in a sequence of n trials there are exactly r E_1. Then by Eq. (2-133)

$$f(r) = P\{X = r\} = \binom{n}{r} p^r q^{n-r}$$

$$F(x) = P\{X \leq x\} = \sum_{r=0}^{[x]} \binom{n}{r} p^r q^{n-r}$$

(2-134)

The distribution function of the random variable X is a step function of the type shown in Fig. 2-23*b*. The corresponding probability density function is shown in Fig. 2-24.

Example 2-31. What is the probability of getting exactly three 1's in five throws of a die? What is the probability of obtaining at most two 1's?

Solution. According to Eq. (2-134), for p = 1/6, q = 5/6, n = 5, and r = 3, one writes

$$P\{X = r = 3\} = \binom{5}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^2 = \frac{250}{7{,}776}$$

For the second part of the problem,

$$F(x) = P\{X \leq 2\} = \binom{5}{0}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^5 + \binom{5}{1}\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^4 + \binom{5}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^3 = 0.96$$

Example 2-32. In a game of n throws of a die, for what value of n is the probability of getting at least two 6's larger than ½?

$$P\{2,3, \ldots ,n \text{ 6's}\} > \tfrac{1}{2}$$

Solution

$$\binom{n}{0}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^n + \binom{n}{1}\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^{n-1} < \frac{1}{2}$$

The numerical answer to this inequality is found to be

$$n \geq 10$$

2-19. Poisson's Distribution.

A random variable X is said to have a Poisson probability distribution if

$$P\{X = x\} = e^{-\lambda}\frac{\lambda^x}{x!}$$

(2-135)

where $\lambda > 0$, $x = 0, 1, 2, \ldots$, and $0! = 1$.

The corresponding cumulative distribution function (CDF) is

$$F(x) = \sum_{k=0}^{[x]} e^{-\lambda}\frac{\lambda^k}{k!} \qquad x \geq 0$$
$$F(x) = 0 \qquad x < 0$$

(2-136)

It is to be noted that $F(x)$ satisfies the conditions required for a distribution function. In fact, $F(x)$ is monotonic, increasing, and, moreover,

$$F(0) = e^{-\lambda}$$
$$F(\infty) = \sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\left(1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \cdots\right) = 1$$

It is of interest to note that the Poisson distribution is a certain type of limiting case of the binomial distribution, in which p is a specified function of n, namely p_n, where

Then

$$\lim_{n\to\infty} np_n = \lambda > 0$$

$$\lim_{n\to\infty} \binom{n}{x} p_n^x (1 - p_n)^{n-x} = e^{-\lambda} \frac{\lambda^x}{x!}$$

(2-137)

The validity of Eq. (2-137) can be checked through the following algebraic manipulations:

$$f(x) = \binom{n}{x} p_n^x (1 - p_n)^{n-x} = \frac{n(n-1)\cdots(n-x+1)}{n^x} \frac{\lambda^x}{x!} (1 - p_n)^{n-x}$$

(2-138)

$$f(x) = \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\cdots\left(1 - \frac{x-1}{n}\right)\frac{\lambda^x}{x!}(1 - p_n)^{n-x}$$

(2-139)

$$f(x) = \frac{(1 - 1/n)(1 - 2/n)\cdots[1 - (x-1)/n]}{(1 - p_n)^x} \frac{\lambda^x}{x!} (1 - p_n)^n$$

(2-140)

But

$$\lim_{n\to\infty} \frac{(1 - 1/n)(1 - 2/n)\cdots[1 - (x-1)/n]}{(1 - p_n)^x} = 1$$

(2-141)

Therefore,

$$\lim_{n\to\infty} (1 - p_n)^n = \lim_{n\to\infty} [(1 - p_n)^{-1/p_n}]^{-\lambda} = e^{-\lambda}$$

(2-142)

Finally, for the limiting case we find

$$f(x) = e^{-\lambda} \frac{\lambda^x}{x!}$$

(2-143)

Thus, in the binomial case, if the number of trials n becomes reasonably large and the probability of individual success p is relatively small, so that their product $np = \lambda$ is of moderate magnitude, the probability of the number of successes in n trials approaches the Poisson distribution. The following relative magnitudes illustrate a common range of application for Poisson's distribution:

$$n > 50 \quad p < 0.1 \quad \lambda < 10$$

In Chap. 6 it will be shown that λ is the "average" value for a random variable with a Poisson distribution.

Example 2-33. Assuming that, on an average, 3 per cent of the output of a factory making certain parts is defective and that 300 units are in a package, what is the probability that, at most, five defective parts may be found in a package?

Solution. The "average" number of defective parts in a package is 300 × 0.03 = 9. Assume a Poisson distribution with this average, i.e.,

$$\lambda = np = 9$$

According to Eq. (2-143), the probability of a box containing x defective parts is

$$\frac{9^x e^{-9}}{x!}$$

$$F(x) = P(X \leq x) = \sum_{k=0}^{[x]} \frac{e^{-9}9^k}{k!}$$

$$F(5) = e^{-9}\left(1 + \frac{9}{1!} + \frac{9^2}{2!} + \frac{9^3}{3!} + \frac{9^4}{4!} + \frac{9^5}{5!}\right)$$

Example 2-34. An industrial process has been running in control with 0.5 per cent defectives. Find the smallest integer k such that the probability of getting k or more defectives in a random sample of 100 is less than 0.10.

Solution. Assuming a Poisson distribution with $p = 0.005$ and $n = 100$, one finds $\lambda = np = 0.5$. Thus it is reasonable to use a Poisson distribution. In this case,

$$P(X \geq k) \leq 0.10$$

$$P(X < k - 1) \geq 0.90$$

$$\sum_{k=1}^{k} \frac{e^{-\lambda}\lambda^{k-1}}{(k-1)!} \geq 0.90$$

$$\sum_{k=1}^{k} \frac{e^{-0.5}0.5^{k-1}}{(k-1)!} \geq 0.90$$

From a Poisson distribution table one finds that

$$k - 1 = 1 \; k = 2$$

2-20. Expected Value of a Random Variable.

Consider a discrete single-variate random variable X and its associated probability function:

$$\begin{matrix}[x_1, x_2, \ldots, x_n] \\ [p_1, p_2, \ldots, p_n]\end{matrix}$$

If the random experiment under consideration is repeated a large number of times, the average or mean value of the numerical function X is found to be

$$\text{Average of } X = \bar{X} = \sum_{k=1}^{n} p_k x_k \tag{2-144}$$

For example, for the experiment of rolling an honest die, one finds

$$X = 1 \cdot \tfrac{1}{6} + 2 \cdot \tfrac{1}{6} + 3 \cdot \tfrac{1}{6} + 4 \cdot \tfrac{1}{6} + 5 \cdot \tfrac{1}{6} + 6 \cdot \tfrac{1}{6} = 3\tfrac{1}{2}$$

More generally, if $\psi(X)$ is a function of a random variable X (also called *a weighting function)*, the mean value of $\psi(X)$ is defined as

$$\text{Mean of } \psi(X) = \overline{\psi(X)} = \sum_{k=1}^{n} p_k \psi(x_k) \tag{2-145}$$

In the literature of probability, the mean of a function is generally referred to as its *expected value.* An alternative notation for denoting the mean value of a

random quantity is a capital E in front of that quantity, for instance, $E(x)$ or $E(X + Y)$ or $E(2X + X^3)$. When the function (X) is of the form $(X) = X^j$, where j is a positive integer, its expected value is called the moment of the jth order of X. For example,

$$E(X) = \bar{X} = \text{first-order moment of } X = \sum_{k=1}^{n} p_k x_k$$

$$E(X^2) = \overline{X^2} = \text{second-order moment of } X = \sum_{k=1}^{n} p_k x_k^2$$

$$E(X^3) = \overline{X^3} = \text{third-order moment of } X = \sum_{k=1}^{n} p_k x_k^3$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

(2-146)

The physical significance of moments is not discussed here. At present the reader is required only to acquaint himself with the concept of Eq. (2-145), that is, how the means of different weighting functions can be calculated. The concept of averaging is of considerable importance in engineering problems. For example, assume that X is a random voltage applied as the input to a device with an input-output relationship

$$Y = \quad (X)$$

Then $E(Y)$ is the d-c level for the output of the system. Similarly, if Y is applied across a unit resistor, the power consumed in the resistor, measured with respect to its d-c level, will have the same numerical value as the second moment of the random variable $(Y - \bar{Y})$, that is, the expectation of

$$\{\psi(X) - E[\psi(X)]\}^2$$

(2-147)

There are at least three special weighting functions of particular interest in probability and information theory. These are

$$X^j \qquad j = 1, 2, 3, \ldots$$
$$e^X \qquad e = \text{base of natural logarithm}$$
$$\log X$$

Without discussing the details at this time, we merely point out the most important application feature of each of the above functions:

$E(X^j)$	This gives moments of different orders of X.
$E(e^X)$	When this mean is known, one can find the values of different moments without recourse to direct computation.
$E(-\log X)$	In the following chapter it will be shown that, when X is taken to be the probability function $f(x)$, the new random variable $[-\log f(x)]$ presents the amount of uncertainty associated with the occurrence of each outcome of the discrete experiment. Therefore, its mean value will stand for the average uncertainty of the system under consideration.

The concept of averaging can be generalized in a direct manner to weighting functions of n random variables associated with an n variate. For example, in the case of a bivariate random variable $[X,Y]$ and a weighting function (X,Y), we have

$$E[\psi(X,Y)] = \sum_j \sum_k \psi(x_k, y_j) p_{kj}$$

(2-148)

PROBLEMS

2-1. Determine whether or not the following relations are correct (the primes denote the complements):

(a) $(A + B)(A + C) = A + BC$

(b) $(A + B) - B = AB'$

(c) $A \quad B = A + B$

(d) $(A - AB)C = A(B + C)'$

(e) $(A + B \quad)C = A' \quad B \quad C$

(f) $(A + B)(B + C)(C + A) = AB + AC + BC$

(g) $(A \cap B) \cap (B' \quad \cap C) = \varnothing$

2-2. Let A, B, C be three arbitrary events of a sample space. Find the expressions for the following cases:

(a) At least one of the three events occurs.

(b) B occurs and either A or C occurs, but not both.

(c) Not more than two occur simultaneously.

2-3. Consider the set of points $S = \{ (X,Y) \}$ shown in Fig. P2-3.

(a) Find the subset a = $\{(x,y) \mid x^2 + y^2 \leq 4\}$.

(b) Describe the subset $b = \{(x,y) \mid y \leq x^2\}$.

(c) Describe the subset $c = \{(x,y) \mid x \leq y^2\}$.

(d) Describe the subset $b \cap c$.

(e) Describe the subset $(b \cup a)c'$.

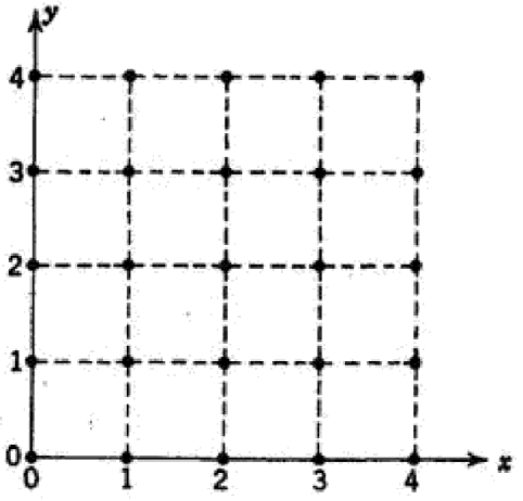


FIG. P2-3

2-4. Given a set $S = \{0,1,2,3,4,5,6,7,8,9,10\}$,

(*a*) Define the function $F_1(x) = x/2$ over S and draw its graph.

(*b*) Define the function $F_2(x) = x + 3$ over S and draw its graph.

(*c*) Determine the subset $a = \{x|(x/2)(x + 3) \leq 4\}$.

2-5. Show the following identities and draw the corresponding Sheffer-stroke diagrams.

(*a*) $(X|(Y|Y)) = X' \cup Y$.

(*b*) $(X|(X|X)) = U$.

(*c*) Verify the identity of the expression for the output as given in

Fig. P2-5
a and b.

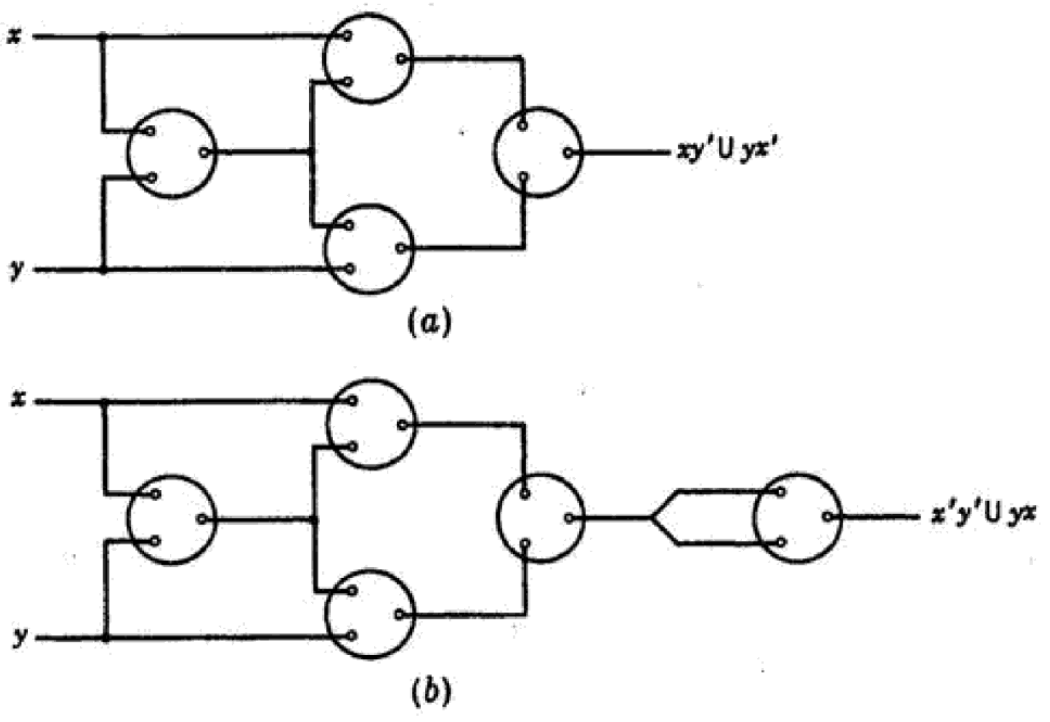


FIG. P2-5

2-6. If $A_1, A_2, \ldots, A_n$ are independent events, show that

$$P\{A_1 + A_2 + \cdots + A_n\} = 1 - P\{A_1'\}P\{A_2'\} \cdots P\{A_n'\}$$

2-7. Two cards are drawn at random successively, the first being replaced before the second is drawn. What is the probability of the first being a club and the second not a queen?

2-8. Two dice are thrown. Denote by A the event that the sum of the faces is even and by B the event that their difference is even. Describe the events $A + B$, AB, $A\ \ B$, AB, and $A\ \ + B$ and find their probabilities.

2-9. Given five letters *a, b, c, d, e*, in how many different ways can one write three-letter words without repeating any letter (*a*) irrespective of their order and (*b*) considering the order of letters?

2-10. In how many different ways can a committee of four men and two women be selected from a total of 20 men and 10 women?

2-11. A survey of 1,000 people has indicated the following results: 714 listen to radio station *A*, 640 to station B, and 850 to station *C*. It also indicated that 530 listen to both *A* and *B*, 375 to both *C* and *B*, and 720 to *A* and *C*. Determine whether these data are not self-contradictory.

2-12. What is the probability of obtaining 8, 9, or 10 with two dice in one trial?

2-13. Two dice are thrown. Let *A* be the event that the sum of the faces is odd and B the event that at least one is a 1. Describe the events AB, $A + B$, AB and find their probabilities.

2-14. What is the probability of drawing a club *or* a face card of any color in a single draw from an ordinary deck of cards?

2-15. Two events *A* and *B* associated with an experiment have respective probabilities of occurrence *p* and *q*. Show that in *n* trials the probability that *AB* occurs K_1 times; AB , K_2 times; A B, K_3 times; and A B , K_4 times is

$$\frac{n!}{K_1!K_2!K_3!K_4!}p^{K_1+K_3}q^{K_2+K_4}(1-p)^{K_2+K_4}(1-q)^{K_3+K_4}$$

2-16. Urn *A* contains seven silver dollars and one $10 gold coin. Urn *B* contains 10 silver dollars. Nine coins are taken from *B* and put in *A;* then eight coins are selected at random from the 17 coins in *A* and put back in urn *B.* If you were to select one of the two urns, which one should you select?

2-17. If the probability of a safe return from a certain trip is $P = 0.9$, what is the probability of exactly four safe returns out of six such trips?

2-18. A single card is removed from a regular deck of cards. From the remainder, we draw two cards and observe that they are both diamonds. What is the probability that the removed card was also a diamond?

2-19. Show that the two relay circuits of Fig. P2-19 are equivalent.

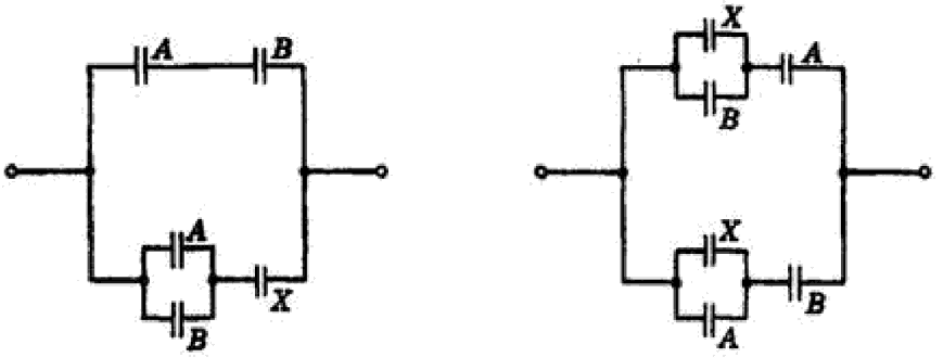


FIG. P2-19

2-20. Express the event of the functioning of the network in FIG. P2-20 in terms of the subevents $E_1, E_2, \ldots, E_6$, where E_k implies the functioning of the kth relay.

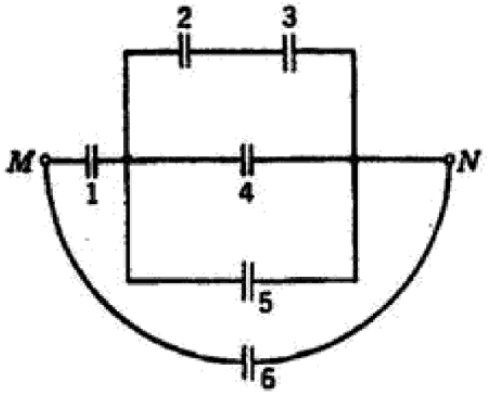


FIG. P2-20

2-21. Two persons toss a coin n times each. What is the probability that they score the same number of heads?

2-22. If a box contains 40 good and 10 defective objects, what is the probability that 10 objects selected at random from the box are all good?

2-23. What is the probability that in a bridge hand a player and his partner have a total of three aces?

2-24. Assuming that the ratio of male to female children is 1:2, find the probability that in a family of six children

(*a*) All children will be of the same sex.

(*b*) The four oldest children will be boys and the two youngest will be girls.

(*c*) Exactly half the children will be boys.

2-25. In a game of bridge, if a player has no ace, what is the probability that his partner has no ace either?

2-26. Find the probability that three, and only three, tails are obtained in a sequence of four tosses of a coin.

2-27. Assuming that the probability of each relay being closed is p, derive the probability for the flow of a current between nodes *A* and *B* of Fig. P2-27.

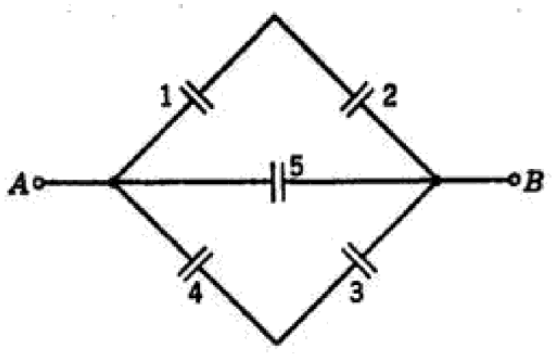


FIG. P2-27

2-28. Same question as in the preceding problem for the networks of Fig. P2-28.

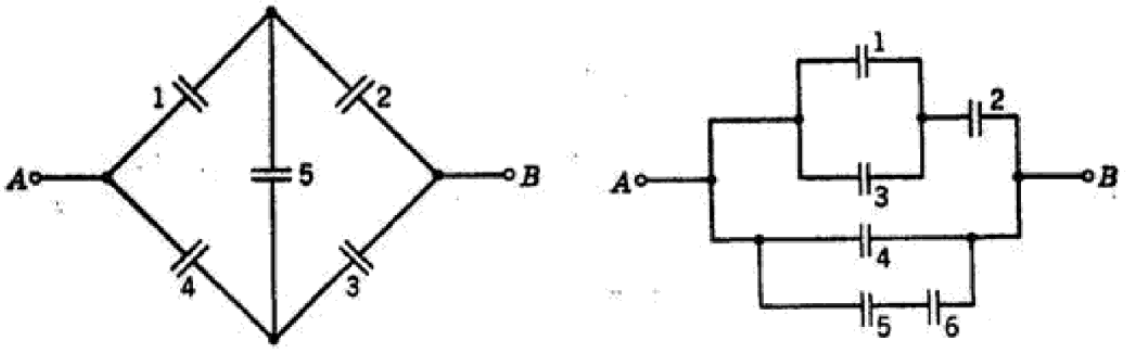


FIG. P2-28

2-29. The following joint probability matrix is given for discrete random variables X and Y. Evaluate the marginal and the conditional probability functions.

$$\begin{bmatrix} \frac{1}{12} & \frac{1}{6} & 0 \\ 0 & \frac{1}{6} & \frac{1}{5} \\ \frac{1}{18} & \frac{1}{4} & \frac{2}{15} \end{bmatrix}$$

2-30. The joint density function for two random variables X and Y is given below:

$f(x,y) = k^2xy2^{-(x+y)}$ for x and y positive integers

$f(x,y) = 0$ elsewhere

Find

$$P\{X < 4, Y < 4\}$$

2-31. Evaluate the probability of getting a four 0, 1, 2, 3, 4, and 5 consecutive times on five throws of a die.

2-32. If the probability of hitting a target is ¼ in each shot, independent of the number of shots fired,

(*a*) What is the probability of the target being hit twice in five shots?

(*b*) What is the probability of the target being hit at least twice in five shots?

2-33. A book of 200 pages contains 100 misprints. Assuming that these are distributed at random, estimate the chances that a page contains at least two misprints.

2-34. The random variable X assumes the values [0,1,2] with respective probabilities [1/3,¼,5/12]. The random variable Y assumes the values [0,1] witch probabilities {¼,¾]. Assuming that the two variables are independent, determine their joint probability functions.

2-35. Study the different probability functions (joint, marginal, and conditional) associated with the following experiment. We draw five cards from an ordinary deck of cards and study the two random variables below:

X, number of aces drawn
Y, number of queens drawn

2-36. A random event E has the probability of occurrence $1/K$ in each experiment independently of the preceding outcome. Determine the following probabilities:

(*a*) E does not occur in n consecutive trials.

(*b*) E occurs in the nth experiment only but not in any of the previous ones.

(*c*) E occurs exactly twice in n experiments.

(*d*) Let $K = 4$, $n = 4$ and evaluate the results of parts (*a*), (*b*), and (*c*).

2-37. *Smith-Jones-Robinson Problem.* The following problem has appeared in the *Scientific American* (vol. 200, no. 2, p. 136, February, 1959) in an entertaining article entitled "Brain-teasers" That Involve Formal Logic.

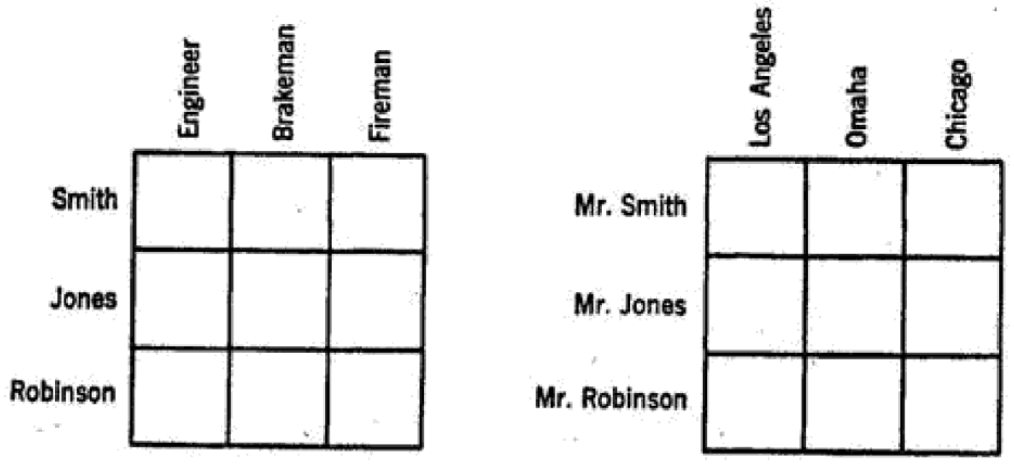


FIG. P2-37

1. Smith, Jones, and Robinson are the engineer, brakeman, and fireman on a train, but not necessarily in that order. Riding the train are three passengers with the same three surnames, to be identified in the following premises by "Mr." before their names.

2. Mr. Robinson lives in Los Angeles.

3. The brakeman lives in Omaha.

4. Mr. Jones long ago forgot all the algebra he learned in high school.

5. The passenger whose name is the same as the brakeman's lives in Chicago.

6. The brakeman and one of the passengers, a distinguished mathematical physicist, attend the same church.

7. Smith beat the fireman at billiards.

Who is the engineer?

HINT: The solution by methods of set theory may become somewhat cumbersome. It is suggested in the above reference to use two matrices as notational aid. Each cell is the intersection of two sets, corresponding to the set of elements contained in the pertinent column and row. Put a 1 or a 0 in a cell indicating that such an intersection is a valid premise or not.

2-38. *Eddington's Controversy.* The following problem exemplifies the type of confusion that existed in probability prior to the introduction of set-theory considerations.

If *A, B, C, D* each speak the truth once in three times (independently), and *A* affirms that *B* denies that *C* declares that *D* is a liar, what is the probability that *D* was speaking the truth?

The following comments on Eddington's problem are given in an article entitled "Brain-Teasers" That Involve Formal Logic by M. Gardner (*op.* cit.).

"Eddington's answer of $^{25}/_{71}$ was greeted by howls of protest from his readers, touching off an amusing controversy that was never decisively resolved. The English astronomer Herbert Dingle, reviewing Eddington's book in *Nature* (Mar. 23, 1935), dismissed the problem as meaningless and symptomatic of Eddington's confused thinking about probability. Theodore Sterne, an American physicist, replied (*Nature*, June 29, 1935) that the problem was not meaningless but lacked sufficient data for a solution. Dingle responded (*Nature*, Sept. 14, 1935) by contending that, if one granted Sterne's approach, there were enough data to reach a solution of exactly $^{1}/_{3}$. Eddington then reentered the fray with a paper entitled The Problem of *A, B, C* and *D* (*Math. Gaz.*, October, 1935), in which he explained in detail how he had calculated his answer."

The difficulty lies chiefly in deciding exactly how to interpret Eddington's statement of the problem. If B is truthful in making his denial, are we justified in assuming that *C* said that *D* spoke the truth? Eddington thought not. Similarly, if *A* is lying, can we then be sure that *B* and *C* said anything at all? Fortunately we can side-step all these verbal difficulties by making (as Eddington did not) the following assumptions: (1) All four men made statements. (2) *A, B*, and *C* each made a statement that either affirmed or denied the statement that follows. (3) A lying affirmation is taken to be a denial, and a lying denial is taken to be an affirmation.

2-39.[19] If a stick is broken at random into three pieces, what is the probability that the pieces can be put together in a triangle?

HINT: The problem, despite its apparently clear statement, is ambiguous. It

requires some additional information about the exact method of breaking the stick. The following two explanations are given in the cited reference.

"One method is to select, independently and at random, two points from the points that range uniformly along the stick, then break the stick at these two points. If this is the procedure to be followed, the answer is ¼, and there is an elegant way of demonstrating it with a geometrical diagram. . . .

"Suppose, however, that we interpret in a different way the statement 'break a stick at random into three pieces.' We break the stick at random, we select randomly one of the two pieces, and we break that piece at random. What are the chances that the three pieces will form a triangle? If after the first break we choose the smaller piece, no triangle is possible."

The latter interpretation of the problem gives $\frac{1}{6}$ for the required probability.

2-40. The joint probability matrix of two variables is given below. Determine whether they are statistically independent.

$$\begin{array}{c} 3 \\ 2 \\ 1 \\ y/x \end{array} \begin{bmatrix} \frac{1}{12} & \frac{1}{36} & \frac{5}{12} \\ \frac{1}{36} & \frac{1}{18} & \frac{5}{36} \\ \frac{1}{12} & \frac{1}{6} & \frac{5}{12} \end{bmatrix} \begin{array}{c} \\ \\ \\ 1 \quad 2 \quad 3 \end{array}$$

2-41. Two urns contain four white and three black balls and three white and seven black balls, respectively. One urn is selected at random and a ball is drawn from it. What is the probability that this ball is white?

2-42. A Markov chain has the transition probability matrix given below:

$$\begin{bmatrix} 0 & \frac{3}{8} & \frac{1}{3} \\ 0 & \frac{7}{12} & \frac{5}{12} \\ 1 & 0 & 0 \end{bmatrix}$$

The three states are initially selected with probabilities ½,. $\frac{1}{6}, \frac{1}{3}$.

(*a*)What is the probability of reaching state 2 via state 1 in one step?

(*b*) What is the probability of reaching state 2 via 1 in two steps?

(*c*) What is the probability of reaching state 3 in two steps?

2-43. Define the probability function for the number of boys in a family of six children, assuming that both sexes are equiprobable and no multiple birth occurs.

2-44. From the joint probability matrix below,

$$\begin{array}{c} y_3 \\ y_2 \\ y_1 \end{array} \begin{bmatrix} 0 & \frac{5}{36} & \frac{1}{3} \\ \frac{1}{12} & \frac{1}{6} & \frac{1}{18} \\ \frac{1}{36} & \frac{1}{4} & 0 \end{bmatrix} \\ \quad x_1 \quad x_2 \quad x_3$$

compute and tabulate:

(*a*) Marginal probability $P_1\{x_k\}$.

(*b*)Marginal probability $P_2\{y_j\}$.

(*c*) $P\{y_j|x_k\}$.

(*d*) $P\{x_k|y_j\}$.

CHAPTER 3

BASIC CONCEPTS OF INFORMATION THEORY: MEMORYLESS FINITE SCHEMES

The object of this chapter is to present the basic elements of information theory of discrete schemes in a manner parallel to the presentation of the elements of discrete probability theory. Our immediate aim is to develop a *measure* for *information content* of a discrete system. That measure will then be used for evaluating the rate of *transmission of information* in a *communication* system. No effort will be made to expound on the philosophical context of terms such as "information measure" or "communication." In order to grasp a basic understanding of this newly developed scientific field, it seems desirable to confine ourselves to an accurate abstract mathematical model rather than to deal with generalities of a semiphilosophical nature. The following approach is suggested:

We shall consider a discrete random experiment and its associated sample space Ω. Let X be a random variable (a real numerical function) associated with Ω; we know that, say, $E(X)$ has a particular physical meaning in regard to the random experiment. That is, if the experiment is repeated a large number of times, the values of X when averaged will approach $E(x)$. In summary, $E(x)$ has given a certain "physical" indication about the experiment. Similarly, $E(X^n)$ has a certain significance in our studies. Then the question arises, could we search for an indicative number associated with the random experiment such that it provides a "measure" of surprise or unexpectedness of occurrence of outcomes of the experiment? Shannon has suggested that the random variable $-\log P\{E_k\}$ is an indicative relative measure of the occurrence of the event E_k. In particular, he shows that the mean of this function is a good indication of the average uncertainty with respect to all the outcomes of the experiment.

The reader should note that the above terms in quotation marks are used here with their common meaning. Their more accurate technical meaning will be defined later.

3-1. A Measure of Uncertainty.

Consider the sample space Ω of events pertaining to a random experiment. We partition the sample space in a finite number of mutually exclusive events E_k, whose probabilities p_k are assumed to be known (Fig. 3-1). The set of all events under consideration can be designated as a row matrix $[E]$ and the corresponding probabilities as another row matrix $[P]$.

$$[E] = [E_1, E_2, \ldots, E_n] \quad \text{with} \quad \bigcup_{k=1}^{n} E_k = U \tag{3-1}$$

$$[P] = [p_1, p_2, \ldots, p_n] \quad \text{with} \quad \sum_{k=1}^{n} p_k = 1 \tag{3-2}$$

Equations (3-1) and (3-2) contain all the information that we have about the probability space - which is called a *complete finite scheme.* For example, the following matrix represents such a situation :

$$\begin{bmatrix} E \\ P \end{bmatrix} = \begin{bmatrix} E_1 & E_2 & E_3 \\ 0.2 & 0.5 & 0.3 \end{bmatrix}$$

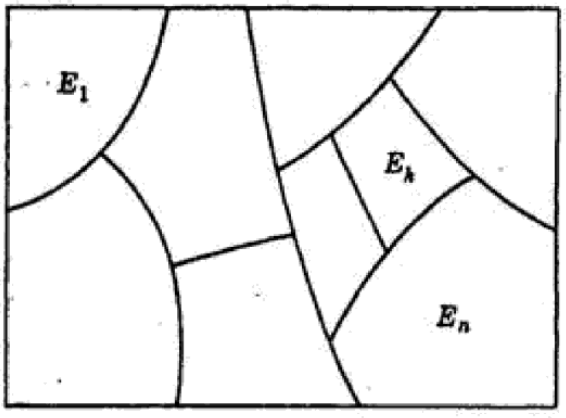


FIG. 3-1. A discrete probability space.

The fundamental problem of interest is to associate a measure of surprise or uncertainty, $H(p_1,p_2, \ldots ,p_n)$, with such probability schemes. Of course at this point it is questionable what is meant by a measure of uncertainty. The clarification of this concept has to come gradually; it is, in essence, the central theme of information theory. The problem can be approached in either of two, not necessarily exclusive, ways:

1. First postulate the desired properties of such an uncertainty measure; then derive the functional form of $H(p_1,p_2, \ldots ,p_n)$. The postulation of the desired properties can be based on some intuitive approach, such as physical motivation or "usefulness" for some purpose, but after such a postulate is adopted, mathematical discipline must prevail and no further intuitive approach may be employed.

2. Assume a known functional $H(p_1,p_2, \ldots ,p_n)$ associated with a finite probability scheme and justify its "usefulness" for the physical problems under consideration.

Our present approach is primarily of type 2. The more mathematically inclined readers who prefer an axiomatic approach are referred to Sec. 3-19 or Feinstein (I).

Shannon and Wiener have suggested the following *measure of uncertainty* or

entropy associated with the sample space of a complete finite scheme.

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i$$

(3-3)[20]

where p_i is the probability of the occurrence of the event E_i as described in Eqs. (3-1) and (3-2). The base of the logarithm is rather arbitrary; however, for communication problems it is convenient to use the binary base.

Our immediate plan in this chapter is first to investigate the principal properties of this suggested measure of uncertainty and to justify its "usefulness" with respect to statistical problems of communication systems. Next we shall generalize this concept to two-dimensional probability schemes, which provide simple models for communications. Finally the discussion will be directed toward more general *n*-dimensional probability schemes. We shall always be restricted to *complete* systems of events; that is, we assume that Eqs. (3-1) and (3-2) are satisfied.

3-2. An Intuitive Justification.

In this section we wish to justify the usefulness of the function suggested in Eq. (3-3) in connection with communication problems. In problems dealing with communication systems, it is often instructive to regard a finite exhaustive probability scheme as a mathematical model for a communication source. In this analogy, any elementary event or outcome, E_k, may be considered as *a letter of the alphabet* of the communication transmitter.

Now consider the random variable

$$X = -\log p$$

(3-4)

defined over the sample space of Fig. 3-1. To each event E_k there corresponds a value x_k of the random variable X, where by hypothesis

$$x_k = -\log P\{E_k\} = -\log p_k$$

(3-5)

The quantity - log p_k is frequently called the amount of *self-information* associated with the event E_k :

$$I(E_k) = -\log p_k$$

(3-6)

The unit of the amount of information is called a *bit*, where one bit is the amount of information associated with the selection of one of two equiprobable (p_k = ½) events. In other words, if the sample space is partitioned into two equally likely events E_1 and *E*2, then

$$I(E_1) = I(E_2) = -\log \tfrac{1}{2} = 1 \text{ bit}$$

(3-7)

A selection between two equally likely events requires one unit of information. If Ω were partitioned into 2^N equally probable events E_k (k = 1, 2, . . . , 2^N), then the self-information associated with any event E_k would be

$$I(E_k) = -\log p_k = -\log 2^{-N} = N \quad \text{bits}$$

(3-8)

The generalization from equiprobable events to the general case is straightforward. In fact, in order to evaluate the self-information associated with a particular event *Eo*, we divide the Ω space in two parts *Eo* and E_0'; thus

$$I(E_0) = -\log p(E_0) = -\log p_0 \qquad \text{bits}$$

(3-9)

For instance, if $p_0 = 1/16$, the occurrence of *Eo* in the average conveys to us 4 bits of information. The measure of self-information is essentially nonnegative:

$$I(E_k) = -\log p_k \geq 0$$

(3-10)

The equality is only by the *certain* event; obviously, no information is conveyed by the knowledge of the occurrence of such an event.

The *average amount of information* or entropy of a finite complete probability scheme is defined by

$$H(X) = \overline{I(E_k)} = -\sum_{k=1}^{n} p_k \log p_k$$

(3-11)

where the random variable *X* is defined over the sample space of events Ω and the events satisfy Eqs. (3-1) and (3-2). *H*(*X*) is the average amount of self-information per event, the average being taken over the entire sample space. In

fact, if $-\log p_k$ indicates the measure of uncertainty associated with the event E_k, then $H(X)$ will clearly represent the mean or the expected value of the uncertainty associated with our probability scheme. As a simple example, let us consider the following three sets of complete events and compare their entropies.

$$\begin{array}{lll} \text{(I)} & E = [A_1, A_2] & P = [1/256, 255/256] \\ \text{(II)} & E = [B_1, B_2] & P = [1/2, 1/2] \\ \text{(III)} & E = [C_1, C_2] & P = [7/16, 9/16] \end{array}$$

The average self-information associated with each of these schemes is given respectively by

$$\begin{array}{ll} \text{(I)} & \bar{I}_1 = -(1/256 \log 1/256 + 255/256 \log 255/256) = 0.0369 \text{ bit} \\ \text{(II)} & \bar{I}_2 = -(1/2 \log 1/2 + 1/2 \log 1/2) = 1 \text{ bit} \\ \text{(III)} & \bar{I}_3 = -(7/16 \log 7/16 + 9/16 \log 9/16) = 0.989 \text{ bit} \end{array}$$

In system I it is relatively easy to guess whether A_1 or A_2 will occur. In system III this guess is much harder, and in II it is most difficult to predict the occurrence of one of the events B_1 or B_2. It is common sense to attribute a larger average uncertainty to system II than to system III and a larger average uncertainty to system III than to system I. This is in agreement with the results obtained by application of the chosen self-information function, that is,

$$\bar{I}_1 < \bar{I}_3 < \bar{I}_2$$

The average uncertainty associated with II is far more than that associated with I. For I, we are almost sure that A_2 generally occurs. For II, the average uncertainty is larger, as it is most difficult to say whether B_1 or B_2 occurs.

3-3. Formal Requirements for the Average Uncertainty.

Shannon's approach, as well as several other authors', in suggesting a suitable H function has been to some extent directed toward an axiomatic description of such functions. The desired H function should have the following basic properties:

1. *Continuity*. That is, if the probabilities of the occurrence of events E_k are slightly changed, the measure of uncertainty associated with the system should vary accordingly in a continuous manner.

$$H(p_1,p_2, \ldots ,p_n) \text{ continuous in } p_k \qquad \begin{matrix} k = 1, 2, \ldots , n \\ 0 \leq p_k \leq 1 \end{matrix}$$

(3-12)

This requirement is obviously in conformity with our physical senses, since a slight change in the probability of the occurrence of an event should not provide us with a significantly large amount of information.

2. *Symmetry*. The H function must be functionally symmetric in every p_k. Indeed, the measure of uncertainty associated with a complete probability set $[E_k, E'_k]$ must be exactly the same as the measure associated with the set $[E_{k'}E_k]$. Our measure must be invariant with respect to the order of these events.

$$H(p_1,p_2, \ldots ,p_n) = H(p_2,p_1, \ldots ,p_n)$$

(3-13)

3. *Extremal Property*. When all the events are equally likely, the average uncertainty must have its largest value. In this case, it is most uncertain which event is going to occur. Conversely, once we know which specific event among a

number of n equally likely events has occurred, we have acquired the largest average amount of information relevant to the occurrence of events of a universe consisting of n complete events.

$$\text{Maximum of } H(p_1,p_2, \ldots ,p_n) = H\left(\frac{1}{n}, \frac{1}{n}, \ldots , \frac{1}{n}\right)$$

(3-14)

4. *Additivity*. Suppose that we have obtained a suitable measure of the average uncertainty $H(p_1,p_2, \ldots ,p_n)$ associated with a complete set of events. Now, let us assume that the event E_n is divided into disjoint subsets (Fig. 3-2) such that

$$E_n = \bigcup_{k=1}^{m} F_k$$

(3-15)

$$p_n = \sum_{k=1}^{m} q_k \qquad P\{F_k\} = q_k$$

(3-16)

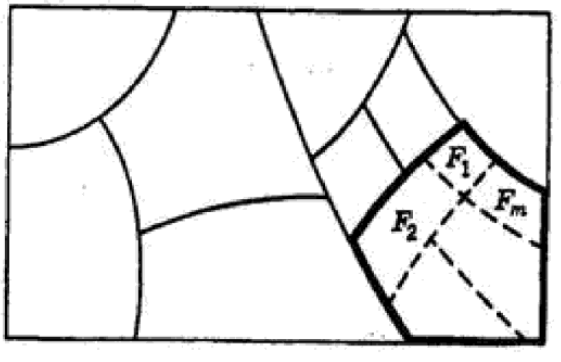


FIG 3-2 A partitioning of the probability apace illustrating the additive property of the information measure

Evidently, the occurrence of the event E_n can be considered as another total sample space where the probabilities associated with events F_k can be normalized in the form

$$\frac{q_1}{p_n} + \frac{q_2}{p_n} + \cdots + \frac{q_m}{p_n} = 1$$

(3-17)

[This recourse provides a rather convenient *relative* frame of reference. That is, we call the event E_n a sample space Ω , associated with the experiments of obtaining all events F_k ($k = 1, 2, \ldots, m$), when we know that E_n is bound to occur.] Therefore we have three probability spaces and hence the following three H functions:

$$\begin{aligned} &H_1(p_1, p_2, \ldots, p_n) \\ &H_2(p_1, p_2, \ldots, p_{n-1}, q_1, q_2, \ldots, q_m) \\ &H_3\left(\frac{q_1}{p_n}, \frac{q_2}{p_n}, \ldots, \frac{q_m}{p_n}\right) \end{aligned}$$

(3-18)

A suitable additive or linear measure which also satisfies our common sense is given by

$$H_2 = H_1 + p_n H_3$$

(3-19)

The occurrence of the weighting factor p_n in this linear form is rather anticipated. However, the uninitiated reader will find the examples of the following section helpful in illustrating this point.

Complying with properties 1 to 4 given above, or with similar requirements, one should be able to derive a functional form for the desired uncertainty function. Such treatments have appeared in the work of Feinstein, Khinchin, Shannon, Schutzenberger, and others. Their findings are not too complicated, but for a detailed presentation much more space is required than is available in the present work. The following references to the literature are recommended for further reading.

1. Fadiev assumes properties 1, 2, and 4 and, subsequent to several lemmas, proves that H must be of the form suggested in Eq. (3-11) except for a multiplicative constant. (See Feinstein [I].)

2. Khinchin assumes properties 1, 3, and 4 and the fact that adding a null set to a complete set of events should not change its entropy, and he derives the form of Eq. (3-11) up to a positive constant multiplier.

3. Schutzenberger [I] aims for a more general axiomatic search for a measure of information associated with a complete set of events. He shows that functions other than the Shannon-Wiener entropy of Eq. (3-11) may also be employed. An example of such a function is given in the work of R. A. Fisher.[21] It should be pointed out, however, that the Shannon-Wiener suggested form is certainly the simplest of all such forms. The present richness and depth of the literature of information theory are to a great extent due to the simplicity of the form of Eq. (3-11).

3-4. *H* Function as a Measure of Uncertainty.

In this section we shall present a treatment concerning the suggested measure of uncertainty. We have discussed that such a measure should obey the following requirements:

$$H(p_1, p_2, \ldots, p_n) \text{ continuous in } p_k \text{ for all } 0 \leq p_k \leq 1$$

(3-20)

$$H(p_k, 1 - p_k) = H(1 - p_k, p_k) \qquad k = 1, 2, \ldots, n$$

(3-21)

$$\text{maximum of } H(p_1,p_2, \ldots ,p_n) = H\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right)$$

(3-22)

$$H(p_1,p_2, \ldots ,p_{n-1},q_1,q_2, \ldots ,q_m) = H(p_1,p_2, \ldots ,p_{n-1},p_n) + p_n H\left(\frac{q_1}{p_n}, \frac{q_2}{p_n}, \ldots, \frac{q_m}{p_n}\right)$$

(3-23)

where

$$p_n = \sum_{k=1}^{m} q_k$$

In the following, we demonstrate that the function defined in Eq. (3-11) satisfies all these requirements.

Property 1: *Continuity*. The entropy function $H(p_1,p_2, \ldots p_n)$ is continuous in each and every independent variable p_k in the interval]0, 1}. The proof follows directly.

$$-H(p_1,p_2, \ldots ,p_n) = p_1 \log p_1 + p_2 \log p_2 + \cdots + p_n \log p_n$$
$$= p_1 \log p_1 + p_2 \log p_2 + \cdots + p_{n-1} \log p_{n-1} + (1 - p_1 - p_2 - \cdots - p_{n-1}) \log (1 - p_1 - p_2 - \cdots - p_{n-1})$$

(3-24)

Note that all independent variables $p_1, p_2, \ldots, p_{n-1}$ and also $(1 - p_1 - p_2 - \cdots - p_{n-1})$ are continuous in]0, 1] and that the logarithm of a continuous function is continuous itself.

Property 2: *Symmetry*. The entropy function is, obviously, a symmetrical function in all variables.

Property 3: *Extremal Value of the Entropy Function*. We should like to show that the entropy function has a maximum when all the individual probabilities are equal.

$$p_1 = p_2 = \cdots = p_n$$

(3-25)

This is in conformity with our intuitive feelings; i.e., in a system where all different states are equiprobable, our average uncertainty will be greatest (in other words, it is most difficult to predict which state is most likely to occur).

We may arbitrarily select p_n as a dependent variable depending on p_k ($k = 1, 2, \ldots, n-1$). In fact,

$$\frac{dH}{dp_k} = \sum_{i=1}^{n} \frac{\partial H}{\partial p_i} \frac{\partial p_i}{\partial p_k} = -\frac{d}{dp_k}(p_k \log p_k) - \frac{d}{dp_n}(p_n \log p_n) \frac{\partial p_n}{\partial p_k}$$

(3-26)

But

$$p_n = 1 - (p_1 + p_2 + \cdots + p_k + \cdots + p_{n-1})$$

(3-27)

Hence

$$\frac{dH}{dp_k} = -(\log_2 e + \log p_k) + (\log_2 e + \log p_n)$$

(3-28)

$$\frac{dH}{dp_k} = -\log \frac{p_k}{p_n}$$

(3-29)

$$\frac{dH}{dp_k} = 0 \qquad \text{yields} \qquad p_k = p_n$$

(3-30)

Since p_k was chosen arbitrarily, we come to the conclusion that, for an extremal point of the H function, we must have

$$p_1 = p_2 = \cdots = p_n = \frac{1}{n}$$

(3-31)

It remains to be shown if the latter relation makes the H function a maximum

and not a minimum. For this we note that

$$H(1,0,0, \ldots ,0) = 0$$

(3-32)

But

$$H\left(\frac{1}{n}, \frac{1}{n}, \ldots , \frac{1}{n}\right) = \log n > 0$$

(3-33)

Thus when all the mutually exclusive events are equiprobable, the H function reaches its maximum value.

Property 4: *Additivity*. We prove the validity of this property by reducing the left member to a form identical with the right member of Eq. (3-23) :

$$\begin{aligned} H(p_1,p_2, \ldots ,p_{n-1},q_1,q_2, \ldots ,q_m) & \\ &= -\sum_{k=1}^{n-1} p_k \log p_k - \sum_{k=1}^{m} q_k \log q_k \\ &= -\sum_{k=1}^{n} p_k \log p_k + p_n \log p_n - \sum_{k=1}^{m} q_k \log q_k \\ &= H(p_1,p_2, \ldots ,p_n) + p_n \log p_n - \sum_{k=1}^{m} q_k \log q_k \end{aligned}$$

(3-34)

But

$$p_n \log p_n - \sum_{k=1}^{m} q_k \log q_k = p_n \sum_{k=1}^{m} \frac{q_k}{p_n} \log p_n - \sum_{k=1}^{m} q_k \log q_k$$
$$= -p_n \sum_{k=1}^{m} \frac{q_k}{p_n} \log \frac{q_k}{p_n}$$
$$= p_n H\left(\frac{q_1}{p_n}, \frac{q_2}{p_n}, \ldots, \frac{q_m}{p_n}\right)$$

(3-35)

This proves the identity of the two sides of Eq. (3-23).

It is to be noted that, since *H* functions are essentially nonnegative, we have

$$H(p_1, p_2, \ldots, p_{n-1}, q_1, q_2, \ldots, q_m) \geq H(p_1, p_2, \ldots, p_{n-1}, p_n)$$

(3-36)

That is, the partitioning of events into subevents cannot decrease the entropy of the system.

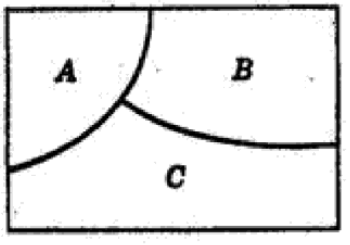


FIG. E3-1

Example 3-1

(*a*) Evaluate the average uncertainty associated with the sample space of

events shown in Fig. E3-1.

$$P\{\ \} = [1/6, 4/15, 8/15]$$

(*b*) Evaluate the average uncertainty pertaining to each of the following probability schemes.

$$[A, M = B \cup C], [B \mid M, C \mid M]$$

(*c*) Verify the rule of the additivity of the entropies.

Solution

(*a*) $H(1/6, 4/15, 8/15) = 1/15(15 \log 5 + 12 \log 3 - 32)$ bits

$H(1/6, 5/6) = 1/15(15 \log 5 - 24)$ bits

(*b*) $H(1/5, 2/5) = 1/15(15 \log 3 - 10)$ bits

(*c*) It is a matter of numerical computation to verify that

$$H(1/6, 4/15, 8/15) = H(1/6, 5/6) + 5/6\, H(1/5, 2/5)$$

Example 3-2. Verify the rule of additivity of entropies for the following probability schemes (Fig. E3-2*a*).

(*a*) $[A,B,C,D]$ (Fig. E3-2*b*).

(*b*) $[A, A\ ]\ [B|A\ , C|A\ , D|A\ ]$ (Fig. E3-2c).

(*c*) (Fig. E3-2*d*.)

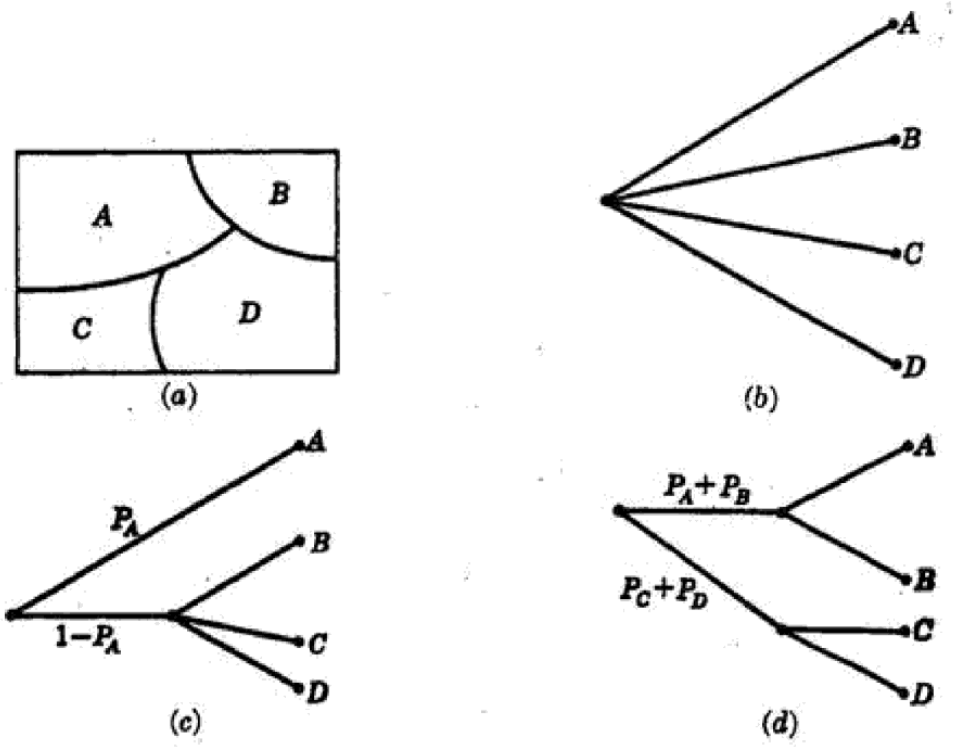


FIG. E3-2

Numerical example:

$$\left[\begin{matrix}\text{Event} \\ \text{Probability}\end{matrix}\right] = \left[\begin{matrix}A & B & C & D \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8}\end{matrix}\right]$$

Solution. The object of the problem is to demonstrate that the average uncertainty in a system is not affected by the arrangement of the events, as long as the probabilities of the individual events do not change.

(*a*)

$$H = -P_A \log P_A - P_B \log P_B - P_C \log P_C - P_D \log P_D$$

where

$$P_A = \tfrac{1}{2}\ P_B = \tfrac{1}{4} P_C = \tfrac{1}{8}\ P_D = \tfrac{1}{8}$$

$$\begin{aligned} H &= -\tfrac{1}{2} \log \tfrac{1}{2} - \tfrac{1}{4} \log \tfrac{1}{4} - \tfrac{1}{8} \log \tfrac{1}{8} - \tfrac{1}{8} \log \tfrac{1}{8} \\ &= \tfrac{1}{2} \log 2 + \tfrac{1}{4} \log 4 + \tfrac{1}{8} \log 8 + \tfrac{1}{8} \log 8 \\ &= \tfrac{1}{2} + \tfrac{1}{2} + \tfrac{3}{8} + \tfrac{3}{8} \\ &= 1\tfrac{3}{4} \text{ bits} \end{aligned}$$

(*b*) According to the additivity property [Eq. (3-19)] of the *H* functions,

$$\begin{aligned} H &= [-P_A \log P_A - (1 - P_A) \log (1 - P_A)] - (1 - P_A) \left(\frac{P_B}{1 - P_A} \log \frac{P_B}{1 - P_A} \right. \\ &\qquad \left. + \frac{P_C}{1 - P_A} \log \frac{P_C}{1 - P_A} + \frac{P_D}{1 - P_A} \log \frac{P_D}{1 - P_A} \right) \\ &= -P_A \log P_A - (1 - P_A) \log (1 - P_A) - P_B \log \frac{P_B}{1 - P_A} \\ &\qquad - P_C \log \frac{P_C}{1 - P_A} - P_D \log \frac{P_D}{1 - P_A} \\ &= -P_A \log P_A - (P_B + P_C + P_D) \log (1 - P_A) - P_B \log P_B + P_B \\ &\quad \log (1 - P_A) - P_C \log P_C + P_C \log (1 - P_A) - P_D \log P_D + P_D \log (1 - P_A) \\ &= -P_A \log P_A - P_B \log P_B - P_C \log P_C - P_D \log P_D \end{aligned}$$

where

$$P_A = \tfrac{1}{2}\ P_B = \tfrac{1}{4}\ P_C = \tfrac{1}{8}\ P_D = \tfrac{1}{8}$$

$$\begin{aligned} H &= -\tfrac{1}{2} \log \tfrac{1}{2} - \tfrac{1}{2} \log \tfrac{1}{2} - \tfrac{1}{2}(\tfrac{1}{2} \log \tfrac{1}{2} + \tfrac{1}{4} \log \tfrac{1}{4} + \tfrac{1}{4} \log \tfrac{1}{4}) \\ &= \tfrac{1}{2} \log 2 + \tfrac{1}{2} \log 2 + \tfrac{1}{2}(\tfrac{1}{2} \log 2 + \tfrac{1}{4} \log 4 + \tfrac{1}{4} \log 4) \\ &= \tfrac{1}{2} + \tfrac{1}{2} + \tfrac{1}{2}(\tfrac{1}{2} + \tfrac{1}{2} + \tfrac{1}{2}) \\ &= 1\tfrac{3}{4} \text{ bits} \end{aligned}$$

(*c*)

$$
\begin{aligned}
H &= -(P_A + P_B)\log(P_A + P_B) - (P_C + P_D)\log(P_C + P_D) \\
&\quad + (P_A + P_B)\left(-\frac{P_A}{P_A + P_B}\log\frac{P_A}{P_A + P_B} - \frac{P_B}{P_A + P_B}\log\frac{P_B}{P_A + P_B}\right) \\
&\quad + (P_C + P_D)\left(-\frac{P_C}{P_C + P_D}\log\frac{P_C}{P_C + P_D} - \frac{P_D}{P_C + P_D}\log\frac{P_D}{P_C + P_D}\right) \\
&= -(P_A + P_B)\log(P_A + P_B) - (P_C + P_D)\log(P_C + P_D) \\
&\quad - P_A\log\frac{P_A}{P_A + P_B} - P_B\log\frac{P_B}{P_A + P_B} - P_C\log\frac{P_C}{P_C + P_D} - P_D\log\frac{P_D}{P_C + P_D} \\
&= -(P_A + P_B)\log(P_A + P_B) - (P_C + P_D)\log(P_C + P_D) \\
&\quad - P_A\log P_A + P_A\log(P_A + P_B) - P_B\log P_B + P_B\log(P_A + P_B) \\
&\quad - P_C\log P_C + P_C\log(P_C + P_D) - P_D\log P_D + P_D\log(P_C + P_D) \\
&= -P_A\log P_A - P_B\log P_B - P_C\log P_C - P_D\log P_D
\end{aligned}
$$

where

$$P_A = \tfrac{1}{2}\ P_B = \tfrac{1}{4}\ Pc = \tfrac{1}{8}\ P_D = \tfrac{1}{8}$$

$$
\begin{aligned}
H &= -(\tfrac{3}{4})\log\tfrac{3}{4} - \tfrac{1}{4}\log\tfrac{1}{4} + \tfrac{3}{4}(-\tfrac{2}{3}\log\tfrac{2}{3} - \tfrac{1}{3}\log\tfrac{1}{3}) \\
&\qquad + \tfrac{1}{4}(-\tfrac{1}{2}\log\tfrac{1}{2} - \tfrac{1}{2}\log\tfrac{1}{2}) \\
&= -\tfrac{3}{4}\log 3 + \tfrac{3}{4}\log 4 + \tfrac{1}{4}\log 4 + \tfrac{3}{4}(-\tfrac{2}{3}\log 2 + \log 3) \\
&\qquad + \tfrac{1}{4}(\tfrac{1}{2}\log 2 + \tfrac{1}{2}\log 2) \\
&= -\tfrac{3}{4}\log 3 + \tfrac{3}{2} + \tfrac{1}{2} + \tfrac{3}{4}(-\tfrac{2}{3} + \log 3) + \tfrac{1}{4}(\tfrac{1}{2} + \tfrac{1}{2}) \\
&= -\tfrac{3}{4}\log 3 + \tfrac{4}{2} - \tfrac{1}{2} + \tfrac{3}{4}\log 3 + \tfrac{1}{4} \\
&= 1\tfrac{3}{4}\ \text{bits}
\end{aligned}
$$

3-5. An Alternative Proof That the. Entropy Function Possesses a Maximum.

The Shannon-Wiener theory of information is strongly linked with the logarithmic function. Thus it is desirable to spend some time investigating some of the basic mathematical properties of the logarithmic function. Such mathematical presentations may seem distant from an immediate engineering application; however, they are of prime significance to those who are interested

in basic research in the field.

First we shall prove a lemma on the convexity of the logarithmic function. Then the lemma will be employed in giving an alternative proof for property 3 of the previous section.

Lemma 1. The logarithmic function is a convex function.

The reader will recall that a function of the real variable $y = f(x)$ is said to be convex upward in a real interval if for any x_1 and x_2 in that interval one has

$$\tfrac{1}{2}[f(x_1) + f(x_2)] \leq f\left(\frac{x_1 + x_2}{2}\right)$$

(3-37)

Geometrically this relation can be simply interpreted by saying that the chord connecting points 1 and 2 lies below the curve. An equivalent definition can be given for a curve that is convex upward in an interval. That is,

$$af(x_1) + (1 - a)f(x_2) \leq f[ax_1 + (1 - a)x_2] \qquad 0 \leq a \leq 1$$

(3-38)

The geometrical interpretation of Eq. (3-38) is that in the interval under consideration the chord lies everywhere below the curve (see Fig. 3-3*a*).

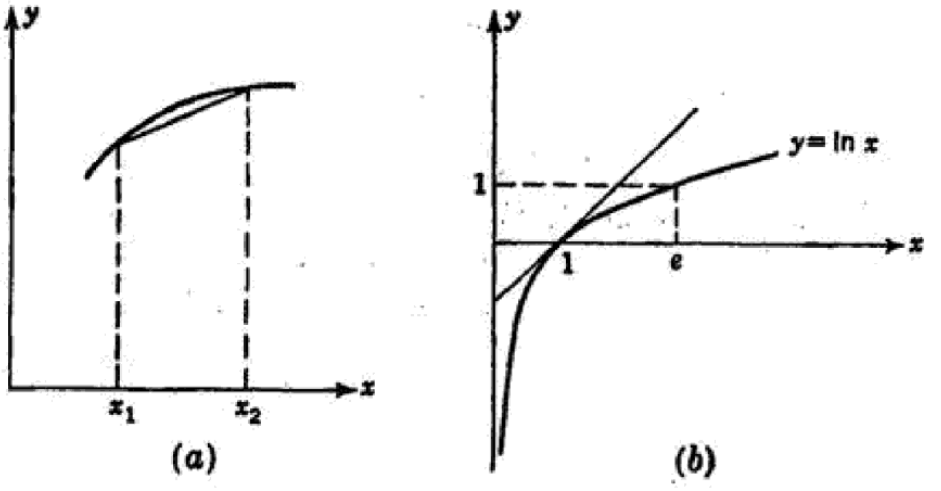


FIG. 3-3. (a) An upward convex function. (b) Logarithmic function is upward convex.

A necessary and sufficient condition for $y = f(x)$ to be convex on the real axis is that

$$\frac{d^2y}{dx^2} \leq 0 \tag{3-39}$$

for every point of the real axis, provided that the second derivative exists. This requirement is satisfied for the function

$$y = \ln x \tag{3-40}$$

In fact,

$$\frac{d^2y}{dx^2} = -\frac{1}{x^2}$$

(3-41)

$$\frac{d^2y}{dx^2} \leq 0 \qquad \text{for } 0 \leq x \leq \infty$$

(3-42)

Note that this property is independent of the base of the logarithm as long as the base is a number greater than unity:

$$\ln x = \ln 2 \cdot \log_2 x$$

(3-43)

Thus we have shown that for positive values of x_1 and x_2

$$\tfrac{1}{2}(\log x_1 + \log x_2) \leq \log \frac{x_1 + x_2}{2}$$

(3-44)

$$(x_1 x_2)^{1/2} \leq \frac{x_1 + x_2}{2}$$

(3-45)

The geometric mean of two positive numbers is smaller than their average.[22]

An alternative formulation of Eq. (3-38) can be given by using the following equivalent criterion for convex functions. [23] If $f(x)$ is convex on the real interval $a \le x \le b$, then for any three values of x, $a \le x_1 \le x_2 \le x_3 \le b$,

$$\begin{vmatrix} x_1 & f(x_1) & 1 \\ x_2 & f(x_2) & 1 \\ x_3 & f(x_3) & 1 \end{vmatrix} \le 0$$

(3-46)

Lemma 2. For any positive number we have

$$\ln x \le x - 1$$

(3-47)

This is a simple conclusion of the convexity of In x. Evidently, the tangent at point $x = 1$ is above the logarithmic curve (Fig. 3-3*b*). The equation of the tangent to the curve at $x = 1$ is given by

$$y_t = \left(\frac{dy}{dx}\bigg|_{x=1} \right) (x - 1)$$

(3-48)

$$y_t = x - 1$$

(3-49)

$$\ln x \le x - 1$$

(3-50)

Again note that this property is equally true for the logarithmic function of the base 2, i.e.,

$$\log x = \ln x \log e \leq (x - 1) \log e$$

The above lemma will be of some use in our future work. At present, we may employ it to give an alternative proof for the fact that the average uncertainty is greatest when all the events are equiprobable. In order to show this, assume that the space of x contains m points, not necessarily with equal probabilities. It is required to show that $H(X)$ is smaller than the entropy of the equiprobable case, that is,

$$H(X) \leq -m\left(\frac{1}{m} \log \frac{1}{m}\right)$$

(3-51)

or to prove

$$H(X) \leq \log m$$

(3-52)

But by definition,

$$H(X) - \log m = \sum_{1}^{m} p_i \log \frac{1}{p_i} + \log \frac{1}{m}$$

(3-53)

Since we are dealing with exhaustive systems, $\log (1/m)$ can be replaced by

$$\left(\log \frac{1}{m}\right)\left(\sum_1^m p_i\right)$$

(3-54)

or

$$H(X) - \log m = \left(\sum_1^m p_i \log \frac{1}{p_i}\right) + \sum_1^m p_i \log \frac{1}{m}$$

(3-55)

$$H(X) - \log m = \sum_1^m p_i \log \frac{1}{p_i m}$$

(3-56)

Applying Lemma 2, we find

$$H(X) - \log m = \sum_1^m p_i \log \frac{1}{p_i m} \leq \sum_1^m p_i \left(\frac{1}{p_i m} - 1\right) \log e$$

(3-57)

$$H(X) - \log m \leq \log_2 e \left[\sum_1^m \left(\frac{1}{m} - p_i \right) \right] = 0$$

(3-58)

$$H(X) \leq \log m$$

(3-59)

The maximum entropy corresponds to the case when all m states have equal probabilities of occurrence $p_i = 1/m$.

3-6. Sources and Binary Sources.

In the study of probability one usually employs concepts of sets but uses certain terminology which differs from that of set theory. Examples of such terminology were given in Sec. 2-6. Similarly, information theory uses certain specialized terms which need to be translated into a more universally understood mathematical form. For our immediate use the following terms are defined:

A source or transmitter is similar to the space of a random experiment. That is, a *source* is the assemblage of all possible events associated with the sample space of a complete random experiment. Each outcome of the experiment corresponds to an elementary output of the source and is called a *symbol or* a *character or* a *letter*.

The finite alphabet of a communication source consists of all its finite distinct characters, much in the same way that the sample space consists of all possible elementary outcomes of a discrete random experiment.

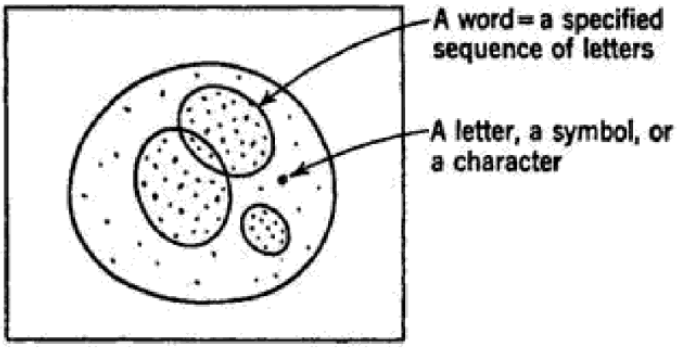


FIG. 3-4. A symbolic illustration of the message space of an independent source; words are specified as sequences of letters (with or without repetition).

A finite sequence of characters may be called a *word* or a *message* in the same way that the sequence of a number of outcomes associated with the repetition of an experiment may be designated as an event. This is schematically illustrated in Fig. 3-4. When the probabilities of the selection of successive letters are independent, we say that the source has no memory. This chapter is devoted to the study of discrete schemes without memory. The study of sources with memory will be deferred until Chap. 11.

A binary source is associated with the sample space of a random binary experiment when the experiment is repeated over and over. In lieu of saying that a random experiment has only two possible exclusive outcomes A and B, we adhere to communication terminology and say that a binary source has an alphabet of two letters A and B. The following three matrices summarize the information-theory performance of a binary source:

$$\text{Alphabet} = \{\text{letters}\} = [A,B]$$
$$\text{Probability matrix } [P] = [p, 1-p] = [p,q]$$
$$\text{Self-information matrix } [I] = [-\log p, -\log (1-p)]$$
$$\text{Average information per letter } H = \bar{I} = -p \log p - (1-p) \log (1-p)$$

The communication entropy for such a system will be

$$H(p) = -p \log p - q \log q = -p \log p - (1 - p) \log (1 - p)$$

(3-61)

A plot of the function $H(p)$ in terms of p is shown in Fig. 3-5. The maximum of this function, as anticipated, occurs at $p = \frac{1}{2}$, for which the entropy becomes 1 bit per letter. If a transmitter is sending the two letters A and B with equal probabilities, the average information per letter is a maximum of 1 bit per letter.

An interesting observation can be made here about the entropy of a binary source. That is, $H(p)$ of Eq. (3-61) is a function concave downward (or convex upward).

$$\tfrac{1}{2}[H(p_1) + H(p_2)] \leq H\left(\frac{p_1 + p_2}{2}\right)$$

(3-62)

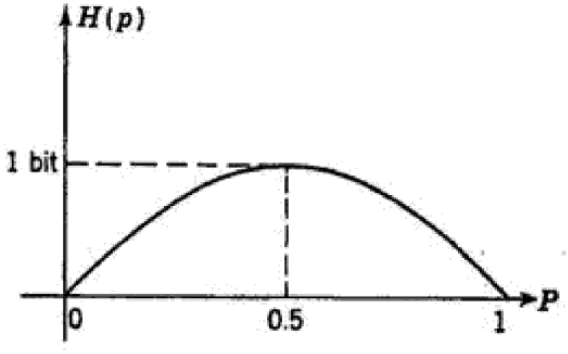


FIG. 3-5. Entropy of an independent binary source.

Suppose that we have three specific binary sources for communication between two stations. If we assume the pertinent probabilities for the first

letters of each source to be p_1, p_2, and $(p_1 + p_2)/2$, the above statement tells us that the average uncertainty of the third source is larger than the mean of the other two. Loosely speaking, it is *relatively* more difficult to predict the transmission of the letters of the third source.

For example, consider the following two binary sources s_1 and s_2.

$$p_{A1} = \tfrac{1}{3} \qquad p_{A2} = \tfrac{1}{4}$$
$$p_{B1} = \tfrac{2}{3} \qquad p_{B2} = \tfrac{3}{4}$$
$$H(s_1) = -\tfrac{1}{3}\log\tfrac{1}{3} - \tfrac{2}{3}\log\tfrac{2}{3} = -\tfrac{2}{3} + \log 3$$
$$H(s_2) = -\tfrac{1}{4}\log\tfrac{1}{4} - \tfrac{3}{4}\log\tfrac{3}{4} = 2 - \tfrac{3}{4}\log 3$$

A third binary source with an average probability $(p_{A1} + p_{A2})/2$ and $(p_{B1} + p_{B2})/2$ per letter will have an average entropy per letter of

$$p_A = \tfrac{1}{2}(\tfrac{1}{3} + \tfrac{1}{4}) = \tfrac{7}{24} \qquad p_B = \tfrac{1}{2}(\tfrac{2}{3} + \tfrac{3}{4}) = \tfrac{17}{24}$$
$$H(s) = -\tfrac{7}{24}\log\tfrac{7}{24} - \tfrac{17}{24}\log\tfrac{17}{24} = 3 + \log 3 - \tfrac{7}{24}\log 7 - \tfrac{17}{24}\log 17$$

The average information per letter for the third source is greater than the mean of the average information associated with letters of the first and the second source.

3-7. Measure of Information for Two-dimensional Discrete Finite Probability Schemes.

In this section, we extend the definition of the measure of information from a one-dimensional to a two-dimensional probability scheme. The content of this section forms an important part of the basic concepts of information theory for several reasons. In the first place, the appropriate generalization from one-dimensional to two-dimensional can be considered as an induction rule for the derivation of the information measure of any finite-dimensional probability space. In the second place, the two-dimensional probability scheme provides the simplest mathematical model for an engineering communication system, that is,

a system with a "transmitter" and a "receiver" or a transducer with in and out ports. Finally the concept of *mutual information*or *transinformation* which forms one of the fundamental concepts of information theory can be discussed in the light of this product space.

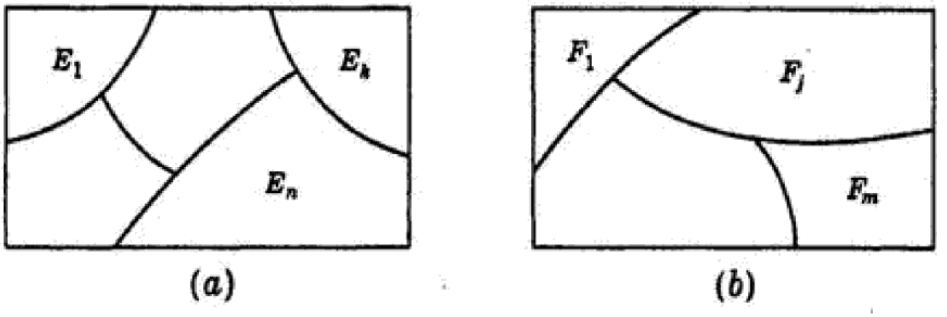


FIG. 3-6. (a) A sample space E. (b) A sample space F.

Consider two finite discrete sample spaces Ω_1, Ω_2, and their product space Ω as illustrated in Figs. 3-6 and 3-7. In Ω_1 and Ω_2 we select complete sets of events in the sense of Eqs. (3-1) and (3-2).

$$\begin{aligned} \{E\} &= [E_1, E_2, \ldots, E_n] \\ \{F\} &= [F_1, F_2, \ldots, F_m] \end{aligned}$$

(3-63)

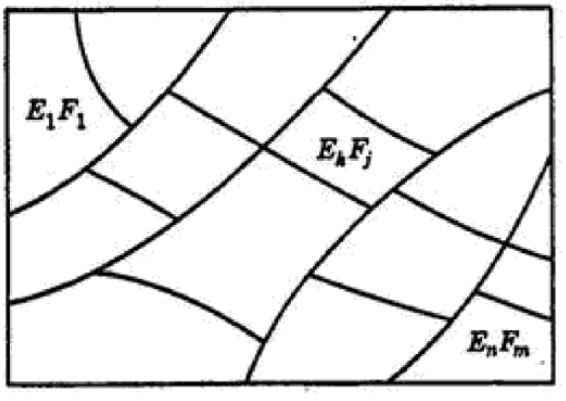


FIG. 3-7. Product space of E ⊗ F.

Each event E_k of Ω_1 may occur in conjunction with any event F_j of Ω_2; thus the following events form a complete set of events in the product space $\Omega_1\Omega_2$.

$$\{EF\} = \begin{bmatrix} E_1F_1 & E_1F_2 & \cdots & E_1F_m \\ E_2F_1 & E_2F_2 & \cdots & E_2F_m \\ \cdots & \cdots & \cdots & \cdots \\ E_nF_1 & E_nF_2 & \cdots & E_nF_m \end{bmatrix}$$

(3-64)

where E_kF_j stands for the simultaneous occurrence of the events E_k and F_j. In this fashion, we are confronted with the following three complete sets of probability schemes:

$$P\{E\} = [P\{E_k\}]$$

(3-65)

$$P\{F\} = [P\{F_j\}]$$

(3-66)

$$P\{EF\} = [P\{E_kF_j\}]$$

(3-67)

No stipulation is made about the independence or dependence of the events E_k

and F_j. Of course, each one of the above three schemes is, by assumption, a finite complete probability scheme. The data pertaining to this fact can be conveniently obtained from the joint probability matrix below.

$$[P\{X,Y\}] = \begin{array}{c} \\ X \end{array} \overset{\begin{array}{cccc} Y & & & \end{array}}{\begin{bmatrix} p\{1,1\} & p\{1,2\} & \cdots & p\{1,m\} \\ p\{2,1\} & p\{2,2\} & \cdots & p\{2,m\} \\ \cdots & \cdots & \cdots & \cdots \\ p\{n,1\} & p\{n,2\} & \cdots & p\{n,m\} \end{bmatrix}}$$

(3-68)

X and Y are random variables, associated with spaces Ω_1 and Ω_2, respectively, and (X,Y) with the product space. The marginal probabilities of the two-dimensional random variables (X,Y) yield the probabilities pertaining to each of the random variables X and Y. For example,

$$\begin{aligned} P\{x_1\} &= P\{E_1\} = P\{E_1F_1 \cup E_1F_2 \cup \cdots \cup E_1F_m\} \\ &= p\{1,1\} + p\{1,2\} + \cdots + p\{1,m\} \end{aligned}$$

(3-69)

$$\begin{aligned} P\{y_2\} &= P\{F_2\} = P\{F_2E_1 \cup F_2E_2 \cup \cdots \cup F_2E_n\} \\ &= p\{1,2\} + p\{2,2\} + \cdots + p\{n,2\} \end{aligned}$$

(3-70)

or

$$P\{x_k\} = \sum_{j=1}^{m} p\{x_k,y_j\}$$

(3-71)

$$P\{y_j\} = \sum_{k=1}^{n} p\{x_k, y_j\}$$

(3-72)

Thus we have three finite complete probability schemes, and naturally there are three corresponding entropies:

$$\begin{aligned} &H(X,Y) = \\ &-p\{1,1\} \log p\{1,1\} - p\{1,2\} \log p\{1,2\} - \cdots - p\{1,m\} \log p\{1,m\} \\ &-p\{2,1\} \log p\{2,1\} - p\{2,2\} \log p\{2,2\} - \cdots - p\{2,m\} \log p\{2,m\} \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &-p\{n,1\} \log p\{n,1\} - p\{n,2\} \log p\{n,2\} - \cdots - p\{n,m\} \log p\{n,m\} \end{aligned}$$

(3-73)

$$\begin{aligned} &H(X) = \\ &-(p\{1,1\} + p\{1,2\} + \cdots + p\{1,m\}) \log (p\{1,1\} + p\{1,2\} + \cdots \\ &\qquad + p\{1,m\}) \\ &-(p\{2,1\} + p\{2,2\} + \cdots + p\{2,m\}) \log (p\{2,1\} + p\{2,2\} + \cdots \\ &\qquad + p\{2,m\}) \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &-(p\{n,1\} + p\{n,2\} + \cdots + p\{n,m\}) \log (p\{n,1\} + p\{n,2\} + \cdots \\ &\qquad + p\{n,m\}) \end{aligned}$$

(3-74)

$$
\begin{aligned}
H(Y) = \\
&-(p\{1,1\} + p\{2,1\} + \cdots + p\{n,1\}) \log (p\{1,1\} + p\{2,1\} + \cdots \\
&\qquad + p\{n,1\}) \\
&-(p\{1,2\} + p\{2,2\} + \cdots + p\{n,2\}) \log (p\{1,2\} + p\{2,2\} + \cdots \\
&\qquad + p\{n,2\}) \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
&-(p\{1,m\} + p\{2,m\} + \cdots + p\{n,m\}) \log (p\{1,m\} + p\{2,m\} + \cdots \\
&\qquad + p\{n,m\})
\end{aligned}
$$

(3-75)

The above three entropies can be expressed in a more condensed fashion by using directly the two-dimensional joint probability matrix of Eq. (3-68):

$$H(X,Y) = -\sum_{k=1}^{k=n}\sum_{j=1}^{j=m} p\{k,j\} \log p\{k,j\}$$

(3-76)

$$H(X) = -\sum_{k=1}^{k=n}\left[\left(\sum_{j=1}^{j=m} p\{k,j\}\right) \log \sum_{j=1}^{j=m} p\{k,j\}\right]$$

(3-77)

$$H(Y) = -\sum_{j=1}^{j=m}\left[\left(\sum_{k=1}^{k=n} p\{k,j\}\right) \log \sum_{k=1}^{k=n} p\{k,j\}\right]$$

(3-78)

$H(X,Y)$ represents the joint entropy, $H(X)$ the marginal entropy of X, and $H(Y)$ the marginal entropy of Y.

The marginal entropies can, of course, be directly expressed in terms of marginal probabilities $p\{x_k\}$ and $p\{y_j\}$, that is,

$$H(X) = -\sum_{k=1}^{k=n} p\{x_k\} \log p\{x_k\}$$

(3-79)

$$H(Y) = -\sum_{j=1}^{j=m} p\{y_j\} \log p\{y_j\}$$

(3-80)

The next section deals with conditional entropies associated with a discrete two-dimensional probability scheme.

3-8. Conditional Entropies.

Reference is made to the matrix of Eq. (3_68) and Fig. 3-7; an event F_j, for example, may occur in conjunction with $E_1, E_2, \ldots,$ or E_n.

$$F_j = \bigcup_{k=1}^{n} E_k F_j$$

(3-81)

$$P\{X = x_k | Y = y_j\} = \frac{P\{X = x_k \cap Y = y_j\}}{P\{Y = y_j\}}$$

(3-82)

or

$$p\{x_k|y_j\} = \frac{p\{k,j\}}{p\{y_j\}}$$

(3-83)

Now consider the following probability scheme:

$$\{E|F_j\} = [E_1|F_j, E_2|F_j, \ldots, E_n|F_j]$$

(3-84)

$$P\{E|F_j\} = \left[\frac{p\{1,j\}}{p\{y_j\}}, \frac{p\{2,j\}}{p\{y_j\}}, \ldots, \frac{p\{n,j\}}{p\{y_j\}}\right]$$

(3-85)

The sum of the elements of this matrix is unity; that is, the probability scheme thus described in not only finite but also complete. Therefore an entropy may be directly associated with such a situation.

$$H(X|y_j) = -\sum_{k=1}^{n} \frac{p\{k,j\}}{p\{y_j\}} \log \frac{p\{k,j\}}{p\{y_j\}}$$

$$= -\sum_{k=1}^{n} p\{x_k|y_j\} \log p\{x_k|y_j\}$$

(3-86)

Now one may take the average of this conditional entropy for all admissible values of y_j, in order to obtain a measure of average conditional entropy of the

system.

$$H(X|Y) = \overline{H(X|y_j)} = \sum_{j=1}^{m} p\{y_j\}[H(X|y_j)]$$
$$= -\sum_{j=1}^{m} p\{y_j\} \sum_{k=1}^{n} p\{x_k|y_j\} \log p\{x_k|y_j\}$$

(3-87)

$$H(X|Y) = -\sum_{j=1}^{m} \sum_{k=1}^{n} p\{y_j\} p\{x_k|y_j\} \log p\{x_k|y_j\}$$

(3-88)

Similarly, one can evaluate the average conditional entropy $H(Y|X)$:

$$H(Y|X) = -\sum_{k=1}^{n} \sum_{j=1}^{m} p\{x_k\} p\{y_j|x_k\} \log p\{y_j|x_k\}$$

(3-89)

The two conditional entropies (the word "average" will be omitted for briefness) can be written as

$$H(X|Y) = -\sum_{j=1}^{m} \sum_{k=1}^{n} p\{x_k, y_j\} \log p\{x_k|y_j\}$$

(3-90)

$$H(Y|X) = -\sum_{k=1}^{n}\sum_{j=1}^{m} p\{x_k,y_j\} \log p\{y_j|x_k\}$$

(3-91)

The conditional entropies along with marginals and the joint entropy compose the five principal entropies pertaining to a joint distribution. All logarithms are taken to the base 2 in order to obtain units in binary digits. Note that all entropies are essentially positive numbers as they are sums of positive numbers.

The physical interpretation of the different entropies will be discussed in the subsequent section.

Example 8-3. Determine five entropies pertaining to the joint probability matrix of Example 2-30.

Solution

$$H(X,Y) = -\sum_{1}^{6}\sum_{1}^{6} P_{ij} \log \tfrac{1}{36} = -\log \tfrac{1}{36} = 2(1 + \log 3)$$

$$H(X) = H(Y) = -\sum_{1}^{6} P_i \log \tfrac{1}{6} = -\log \tfrac{1}{6} = 1 + \log 3$$

$$H(X|Y) = H(Y|X) = -\sum_{1}^{6}\sum_{1}^{6} P_{ij} \log \tfrac{1}{6} = 1 + \log 3$$

3-9. A Sketch of a Communication Network.

In this section, we wish to present an informal sketch of a model for a communication network. In contrast to the material of the previous sections, the content of this section is not presented in a strict mathematical frame. The words source, load, channel, transducer, transmitter, and receiver are used in their common engineering sense. Later on, we shall assign a strict mathematical description to some of these words, but for the present the reader is cautioned

against any identification of these terms with similar terms defined in the professional literature.

In the study of physical systems from a systems engineering point of view, we generally focus our attention on a number of points of entry to the system. For example, in ordinary electric networks, we may be interested in the study of voltage-current relationships at the same port of entry in the network (Fig. 3-8*a*). This is generally known as a one-port system.

When the voltage-current relationships between two ports of entries are of interest, the situation is that of a two-port system. In a two-port system, a physical driving force is applied to one port and its effect observed at a second port. The second port may be connected to a "receiver" or "load" (Fig. 3-8*b*). Such a system is usually known as a two-port, or a loaded transducer. More generally, in many physical problems we may be interested in the study of an *n*-port network (Fig. 3-8*c*). From linear network theory, we know that a complete study of *n*-port systems requires a knowledge of transmission functions between different ports. For example, if we concentrate on different impedances of a network, the following matrices are considered for a general study of a one-port, two-port, and *n*-port, respectively.

$$[Z_{11}] \qquad \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix} \qquad \begin{bmatrix} Z_{11} & Z_{12} & \cdots & Z_{1n} \\ Z_{21} & Z_{22} & \cdots & Z_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ Z_{n1} & Z_{n2} & \cdots & Z_{nn} \end{bmatrix}$$

(3-92)

(The impedances are used in the ordinary circuit sense, Z_{kj} being the transfer impedance between the *k*th and the *j*th port.)

An equivalent interpretation can be made for the study of probabilistic systems. In fact, the systems point of view does not rely on the deterministic or probabilistic description of the performance. It is based on the *ports* of application of stimuli and observation of responses. For instance, consider a

source of communication with a given *alphabet*. The source is linked to the *receiver via* a *channel*. The system may be described by a joint probability matrix, that is, by giving the probability of the joint occurrence of two symbols, one at the input and the other at the output. The joint probability matrix may be designated by

$$[P\{X,Y\}] = \begin{bmatrix} P\{x_1,y_1\} & P\{x_1,y_2\} & \cdots & P\{x_1,y_n\} \\ P\{x_2,y_1\} & P\{x_2,y_2\} & \cdots & P\{x_2,y_n\} \\ \cdots & \cdots & \cdots & \cdots \\ P\{x_m,y_1\} & P\{x_m,y_2\} & \cdots & P\{x_m,y_n\} \end{bmatrix}$$

(3-93)

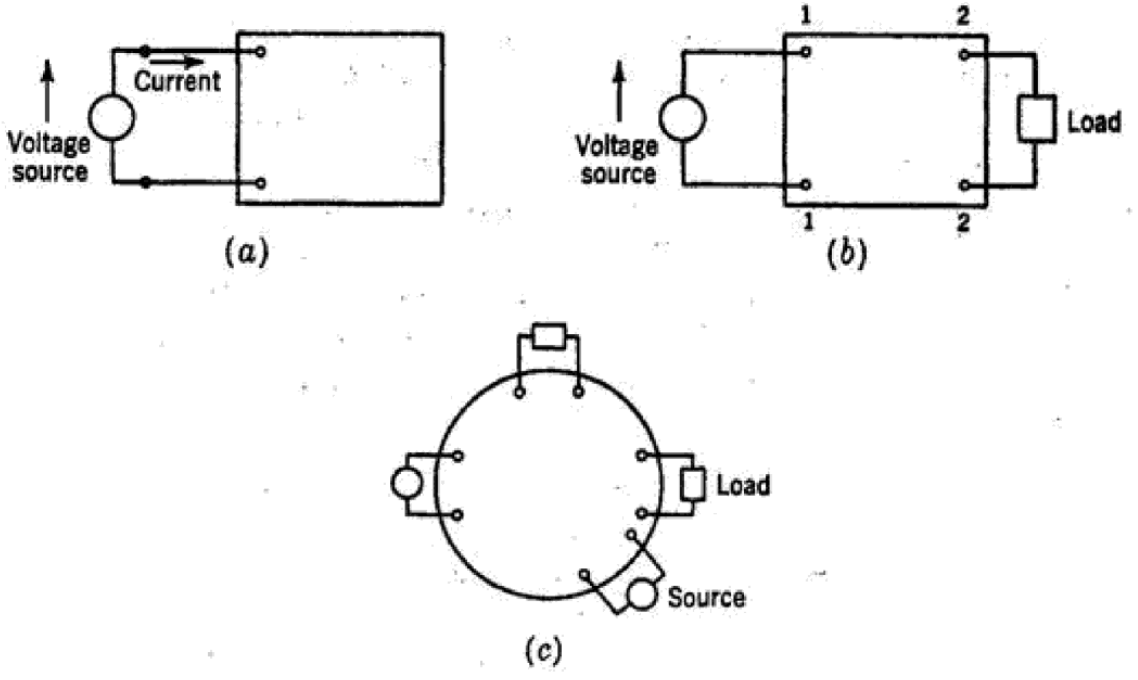


FIG. 3-8. (a) A one-port network. (b) A two-port analog of a channel connecting a source and a receiver. (c) An n-port analog of a communication system consisting of several sources, channels, and sinks.

But in a product space of the two random variables X and Y there are five basic probability schemes of interest. These are

$$[P\{X,Y\}] \qquad \textbf{joint probability matrix} \tag{3-94}$$

$$[P\{X\}] \qquad \textbf{marginal probability matrix of } X \tag{3-95}$$

$$[P\{Y\}] \qquad \textbf{marginal probability matrix of } Y \tag{3-96}$$

$$\begin{aligned} &[P\{X|Y\}] \quad \textbf{conditional probability matrix} \\ &[P\{Y|X\}] \quad \textbf{conditional probability matrix} \end{aligned} \tag{3-97}$$

Thus we are naturally led to five distinct functions in the study of a simple communication model.

This idea can be generalized to n-port communication systems. The problem is similar to the study of an n-dimensional discrete random variable or product space. In each product probability space there are a finite number of basic probability schemes (marginals and conditionals of different orders). With each of these schemes, we may associate an entropy and directly interpret its physical significance.

A source of information is in a way similar to the driving source in a circuit; the receiver is similar to the load, and the channel acts as the network connecting the load to the source. The following interpretations of the different

entropies for a two-port communication system seem pertinent.

$H(X)$ Average information per character at the source, or the entropy of the source.

$H(Y)$ Average information per character at the destination, or the entropy at the receiver.

$H(X,Y)$ Average information per pairs of transmitted and received characters, or the average uncertainty of the communication system as a whole.

$H(Y|X)$ A specific character x_i, being transmitted; one of the permissible y_j may be received with a given probability. The entropy associated with this probability scheme when x_i, covers sets of all transmitted symbols, that is, $\overline{H(Y|x_i)}$, is the conditional entropy $H(Y|X)$, a measure of information about the receiving port, where it is known that x is transmitted.

$H(X|Y)$ A specific character y_j being received; this may be a result of transmission of one of the x_i with a given probability. The entropy associated with this probability scheme when y_j covers all the received symbols, that is, $\overline{H(X|y_j)}$, is the entropy $H(X|Y)$ or equivocation, a measure of information about the source, where it is known that Y is received.

$H(X)$ and $H(Y)$ give indications of the probabilistic nature of the transmission and reception ports, respectively. $H(Y|X)$ gives an indication of the *noise* or *error* in the channel, and $H(X|Y)$ indicates a measure of equivocation, that .is, how well one can recover the input content from the output.

All the probabilities encountered in the two-dimensional case can be derived from the joint probability matrix. Thus, a joint probability matrix specifies a communication channel, in much the same way that an impedance or admittance matrix specifies the performance of an ordinary linear two-port network with respect to its ports.

3-10. Derivation of the Noise Characteristics of a Channel.

In communication problems in general, the joint probability matrix is not given. It is customary to specify the *noise characteristics* of a channel and the source alphabet probabilities. From these data we can directly derive the joint and the output probability matrices. For example, the joint probability matrix is

$$\begin{bmatrix} p\{x_1\}p\{y_1|x_1\} & p\{x_1\}p\{y_2|x_1\} & \cdots & p\{x_1\}p\{y_n|x_1\} \\ p\{x_2\}p\{y_1|x_2\} & p\{x_2\}p\{y_2|x_2\} & \cdots & p\{x_2\}p\{y_n|x_2\} \\ \cdots & \cdots & \cdots & \cdots \\ p\{x_m\}p\{y_1|x_m\} & p\{x_m\}p\{y_2|x_m\} & \cdots & p\{x_m\}p\{y_n|x_m\} \end{bmatrix}$$

which can be written as

$$[P\{X\}][P\{Y|X\}] = [P\{X,Y\}]$$

(In this form we assume that the marginal probability matrix is written in a diagonal form.)

Similarly, if for convenience $[P\{X\}]$ is written in the form of a row matrix, we have

$$[P\{X\}][P\{Y|X\}] = [P\{Y\}]$$

where $[P\{Y\}]$ will also be a row matrix designating the probabilities of the output alphabets.

This section offers for discussion two particularly simple communication channels:

1. Discrete noise-free channel
2. Discrete channel with independent input-output

Discrete Noise-free Channel. In such channels, as their name indicates, every

letter of the input alphabet is in a one-to-one correspondence with a letter of the output alphabet. The joint probability matrix, as well as the channel probability matrix, is of the diagonal form:

$$[P\{X,Y\}] = \begin{bmatrix} p\{x_1,y_1\} & 0 & \cdots & 0 \\ 0 & p\{x_2,y_2\} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & p\{x_n,y_n\} \end{bmatrix}$$

(3-98)

$$[P\{X|Y\}] = [P\{Y|X\}] = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

(3-99)

For a noise-free channel the entropies are

$$H(X,Y) = H(X) = H(Y) = -\sum_{i=1}^{n} p\{x_i,y_i\} \log p\{x_i,y_i\}$$

(3-100)

$$H(Y|X) = H(X|Y) = 0$$

(3-101)

The interpretation of these formulas for a communication system is rather clear. To each transmitted symbol in a noise-free channel there corresponds

one, and only one, received symbol. The average uncertainty at the receiving end is exactly the same as at the sending end. The individual conditional entropies are all equal to zero, a fact that reiterates a nonambiguous or noise-free transmission.

Discrete Channel with Independent Input-Output. In a similar fashion, one can visualize a channel in which there is no correlation between input and output symbols. That is, an input letter x_i can be received as any one of the symbols y_j of the receiving alphabet with equal probability. As will be shown, such a system is a degenerate one as it does not transmit any information. The joint probability matrix has n identical columns.

$$[P\{X,Y\}] = \begin{array}{c} \\ X \end{array}\overset{\begin{array}{cccc} Y & & & \end{array}}{\begin{bmatrix} p & p_1 & \cdots & p_1 \\ p_2 & p_2 & \cdots & p_2 \\ \cdots & \cdots & \cdots & \cdots \\ p_m & p_m & \cdots & p_m \end{bmatrix}} \qquad \sum_i^m p_i = \frac{1}{n}$$

(3-102)

The input and output symbol probabilities are statistically independent of each other, that is,

$$p\{x_i,y_j\} = p_1\{x_i\}p_2\{y_j\}$$

(3-103)

This can be shown directly by calculation:

$$p_{ij} = np_i\left(\sum_1^m p_j\right) = np_i\frac{1}{n} = p_i$$

(3-104)

From this one concludes that

$$p\{x_i|y_j\} = p_1\{x_i\} = np_i$$

(3-105)

$$p\{y_j|x_i\} = p_2\{y_j\} = \frac{1}{n}$$

(3-106)

The different entropies can be computed directly:

$$H(X,Y) = -n\left(\sum_{i=1}^{m} p_i \log p_i\right)$$

(3-107)

$$H(X) = -\sum_{i=1}^{m} np_i \log np_i = -n\left(\sum_{i=1}^{m} p_i \log p_i\right) - \log n$$

(3-108)

$$H(Y) = -n\left(\frac{1}{n}\log\frac{1}{n}\right) = \log n$$

(3-109)

$$H(X|Y) = -\sum_{i=1}^{m} np_i \log np_i = H(X)$$

(3-110)

$$H(Y|X) = -\sum_{i=1}^{m} np_i \log \frac{1}{n} = \log n = H(Y)$$

(3-111)

The interpretation of the above formula is that a channel with independent input and output ports conveys no information whatsoever. To mention a network analogy, this channel seems to have the largest internal "loss," like a resistive network, in contrast to the noise-free channel which resembles a "lossless" network.

3-11. Some Basic Relationships among Different Entropies.

In this section we should like first to investigate some of the fundamental mathematical relations that exist among different entropies in a simple two-port communication system and then point out their significance in communication theories. Our starting point is the evident fact that the different probabilities in a two-dimensional distribution (product space) are interrelated, plus the fact that the chosen logarithmic weighting function is a convex function on the positive real axis. We begin with the basic relationship that exists among the joint, marginal, and conditional probabilities, that is,

$$p\{x_k,y_j\} = p\{x_k|y_j\} \cdot p\{y_j\} = p\{y_j|x_k\} \cdot p\{x_k\}$$

(3-112)

$$\log p\{x_k,y_j\} = \log p\{x_k|y_j\} + \log p\{y_j\}$$
$$= \log p\{y_j|x_k\} + \log p\{x_k\}$$

(3-113)

The direct substitution of these relations in the defining equations of the entropies leads to the following basic identities:

$$H(X,Y) = H(X|Y) + H(Y)$$

(3-114)

$$H(X,Y) = H(Y|X) + H(X)$$

(3-115)

Next we should like to establish a fundamental inequality first shown by Shannon, namely,

$$H(X) \geq H(X|Y)$$

(3-116)

For the proof of this inequality, we employ once again Eq. (3-50) for log $(p\{x_k;/p(x_k|y_j\})$.

$$H(X|Y) - H(X) = \sum_{j=1}^{m}\sum_{k=1}^{n} p\{x_k,y_j\} \log \frac{p\{x_k\}}{p\{x_k|y_j\}}$$

$$\leq \sum_{j=1}^{m}\sum_{k=1}^{n} p\{x_k,y_j\} \left(\frac{p\{x_k\}}{p\{x_k|y_j\}} - 1\right) \log e$$

(3-117)

But the right side of this inequality is identically zero as

$$\sum_{j=1}^{m}\sum_{k=1}^{n} (p\{x_k\} \cdot p\{y_j\} - p\{x_k,y_j\}) \log e = \sum_{j=1}^{m} (p\{y_j\} - p\{y_j\}) \log e = 0$$

(3-118)

Hence,

$$H(X) \geq H(X|Y)$$

(3-119)

and similarly one shows that

$$H(Y) \geq H(Y|X)$$

(3-120)

The equality signs hold if, and only if, X and Y are statistically independent. It is only in such a case that our key inequality Eq. (3-50) becomes an equality (at point $x = 1$), that is,

$$\frac{p\{x_k\}}{p\{x_k|y_j\}} = 1$$

(3-121)

for all permissible values of k and j. This is the case of independence between X and Y.

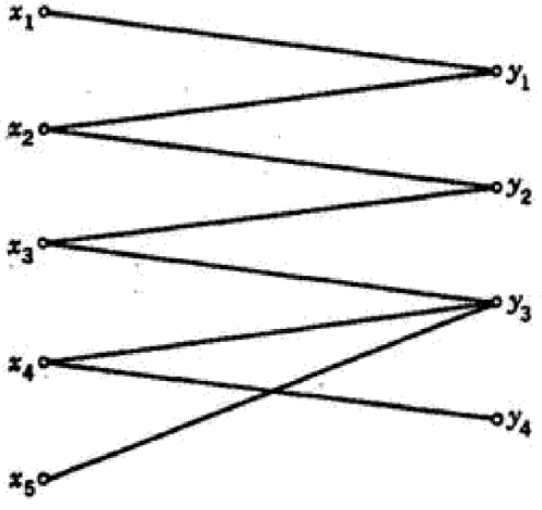


FIG. E3-4

Example 3-4. A transmitter has an alphabet consisting of five letters $\{x_1,x_2,x_3,x_4,x_5\}$ and the receiver has an alphabet of four letters $\{y_1,y_2,y_3,y_4\}$. The joint probabilities for the communication are given below. See Fig. E3-4.

$$
\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{array}
\begin{array}{c}
\begin{array}{cccc} y_1 & y_2 & y_3 & y_4 \end{array} \\
\begin{bmatrix} 0.25 & 0 & 0 & 0 \\ 0.10 & 0.30 & 0 & 0 \\ 0 & 0.05 & 0.10 & 0 \\ 0 & 0 & 0.05 & 0.10 \\ 0 & 0 & 0.05 & 0 \end{bmatrix}
\end{array}
$$

Determine the different entropies for this channel.

Solution

$$
\begin{aligned}
f_1(x_1) &= 0.25 & f_2(y_1) &= 0.25 + 0.10 = 0.35 \\
f_1(x_2) &= 0.10 + 0.30 = 0.40 & f_2(y_2) &= 0.30 + 0.05 = 0.35 \\
f_1(x_3) &= 0.05 + 0.10 = 0.15 & f_2(y_3) &= 0.10 + 0.05 + 0.05 = 0.20 \\
f_1(x_4) &= 0.05 + 0.10 = 0.15 & f_2(y_4) &= 0.10 \\
f_1(x_5) &= 0.05
\end{aligned}
$$

$$
\begin{aligned}
f(x_1|y_1) &= \frac{f(x_1,y_1)}{f_2(y_1)} = \frac{0.25}{0.35} = \frac{5}{7} & f(y_1|x_1) &= \frac{f(x_1,y_1)}{f_1(x_1)} = \frac{0.25}{0.25} \\
f(x_2|y_2) &= \frac{0.30}{0.35} = \frac{6}{7} & f(y_2|x_2) &= \frac{0.30}{0.40} = \frac{3}{4} \\
f(x_3|y_3) &= \frac{0.10}{0.20} = \frac{1}{2} & f(y_3|x_3) &= \frac{0.10}{0.15} = \frac{2}{3} \\
f(x_4|y_4) &= \frac{0.10}{0.10} = 1 & f(y_4|x_4) &= \frac{0.10}{0.15} = \frac{2}{3} \\
f(x_2|y_1) &= \frac{0.10}{0.35} = \frac{2}{7} & f(y_1|x_2) &= \frac{0.10}{0.40} = \frac{1}{4} \\
f(x_3|y_2) &= \frac{0.05}{0.35} = \frac{1}{7} & f(y_2|x_3) &= \frac{0.05}{0.15} = \frac{1}{3} \\
f(x_4|y_3) &= \frac{0.05}{0.20} = \frac{1}{4} & f(y_3|x_4) &= \frac{0.05}{0.15} = \frac{1}{3} \\
f(x_5|y_3) &= \frac{0.05}{0.20} = \frac{1}{4} & f(y_3|x_5) &= \frac{0.05}{0.05} = 1
\end{aligned}
$$

$$
\begin{aligned}
H(X,Y) &= -\sum_x \sum_y f(x,y) \log f(x,y) \\
&= -0.25 \log 0.25 - 0.10 \log 0.10 - 0.30 \log 0.30 - 0.05 \log 0.05 \\
&\quad - 0.10 \log 0.10 - 0.05 \log 0.05 - 0.10 \log 0.10 - 0.05 \log 0.05 \\
&= 2.665 \\
H(X) &= -\sum_x \sum_y f(x,y) \log f_1(x) \\
&= -0.25 \log 0.25 - 0.10 \log 0.40 - 0.30 \log 0.40 - 0.05 \log 0.15 \\
&\quad - 0.10 \log 0.15 - 0.05 \log 0.15 - 0.10 \log 0.15 - 0.05 \log 0.05 \\
&= 2.066 \\
H(Y) &= \sum_x \sum_y f(x,y) \log f_2(y) \\
&= -0.25 \log 0.35 - 0.10 \log 0.35 - 0.30 \log 0.35 - 0.05 \log 0.35
\end{aligned}
$$

$$- 0.10 \log 0.20 - 0.05 \log 0.20 - 0.05 \log 0.20 - 0.10 \log 0.10$$
$$= 1.856$$

$$H(Y|X) = -\sum_x \sum_y f(x,y) \log \frac{f(x,y)}{f_1(x)}$$
$$= -0.10 \log \tfrac{1}{4} - 0.30 \log \tfrac{3}{4} - 0.05 \log \tfrac{1}{8}$$
$$- 0.10 \log \tfrac{3}{8} - 0.05 \log \tfrac{1}{8} - 0.10 \log \tfrac{3}{8}$$
$$= 0.600$$

$$H(X|Y) = -\sum_x \sum_y f(x,y) \log \frac{f(x,y)}{f_2(y)}$$
$$= -0.25 \log \tfrac{5}{7} - 0.10 \log \tfrac{3}{7} - 0.30 \log \tfrac{5}{7} - 0.05 \log \tfrac{1}{7}$$
$$- 0.10 \log \tfrac{1}{2} - 0.05 \log \tfrac{1}{4} - 0.05 \log \tfrac{1}{4}$$
$$= 0.809$$

Note that

$$H(X,Y) < H(X) + H(Y)$$
$$2.665 < 2.066 + 1.856$$

and

$$H(X,Y) = H(Y) + H(X|Y) = H(X) + H(Y|X)$$
$$2.665 = 1.856 + 0.809 = 2.066 + 0.600$$

3-12. A Measure of Mutual Information.

Consider a discrete communication system with given joint probabilities between its input and output terminals. Each transmitted symbol x_i, while going through the channel has a certain probability $P\{y_j|x_i\}$ of being received as a particular symbol y_j. In the light of previous developments, one may look for a function relating a measure of mutual information between x_i and y_j. In other words, how many bits of information do we obtain in knowing that y_j corresponds to x_i when we know the over-all probability of x_i happening along with different y? In order to avoid a complex mathematical presentation, we follow a procedure similar to that of Sec. 3-3. We assume a definition for mutual information and justify its agreement with that of the previously adopted

definition of the entropy. Finally, we shall investigate some of the properties of the suggested measure of mutual information. A measure for the mutual information contained in $(x_i|y_j)$ can be given as

$$I(x_i;y_j) = \log_2 \frac{p\{x_i|y_j\}}{p\{x_i\}} = \log \frac{p\{x_i,y_j\}}{p\{x_i\}p\{y_j\}}$$

(3-122)

This expression gives a reasonable measure of mutual information conveyed by a pair of symbols (x_i,y_j). For a moment, we concentrate on the received symbol y_j. Suppose that an observer is stationed at the receiver end at the position of the signal y_j. His a priori knowledge that a symbol x_i is being transmitted is the marginal probability $p\{x_i\}$, that is, the sum of the probabilities of x_i being transmitted and received as any one of the possible y_j. The a posteriori knowledge of our observer is based on the conditional probability of x_i being transmitted, given that a particular y_j is received, that is, $p\{x_i|y_j\}$. Therefore, loosely speaking, for this observer the gain of information is the logarithm of the ratio of his final and initial ignorance or uncertainties. However, the mathematically inclined reader may wish to forgo such justification and use (3-122) as a definition.

The following elementary properties can be derived for the mutual information function:

1. *Continuity.* $I(x_i;y_j)$ is a continuous function of $p\{x_i|y_j\}$

2. *Symmetry or reciprocity*. The information conveyed by y_j about x_i is the same as the information conveyed by x_i about y_j, that is,

$$I(x_i;y_j) = I(y_j;x_i)$$

(3-123)

Obviously, Eq. (3-122) is symmetric with respect to x_i and y_j.

3. *Mutual and self-information.* The function $I(x_i;x_i)$ may be called the self-information of a symbol x_i. That is, if an observer is stationed at the position of the symbol x_i his a priori knowledge of the situation is that x_i will be transmitted with the probability $p\{x_i\}$ and his a posteriori knowledge is the certainty that x_i; has been transmitted; thus

$$I(x_i) = I(x_i;x_i) = \log \frac{1}{p\{x_i\}}$$

(3-124)

Obviously,

$$I(x_i;y_j) \leq I(x_i;x_i) = I(x_i)$$

(3-125)

$$I(x_i;y_j) \leq I(y_j;y_j) = I(y_j)$$

(3-126)

An interesting interpretation of the concept of mutual information can be given by obtaining the average of the mutual information per symbol pairs, that is,

$$I(X;Y) = \overline{I(x_i;y_j)} = \sum_j \sum_i p\{x_i,y_j\} I(x_i;y_j)$$

(3-127)

$$I(X;Y) = \sum_j \sum_i p\{x_i,y_j\} \log \frac{p\{x_i|y_j\}}{p\{x_i\}}$$

(3-128)

It could be ascertained that this definition provides a proper measure for the mutual information of all the pairs of symbols. On the other hand, the definition ties in with our previously defined basic entropy formulas. Indeed, by direct application of the defining equations one can show that

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

(3-129)

$$I(X;Y) = H(X) - H(X|Y)$$

(3-130)

$$I(X;Y) = H(Y) - H(Y|X)$$

(3-131)

The entropy corresponding to the mutual information, that is, $I(X;Y)$, indicates a measure of the information transmitted through the channel. For this reason it is referred to as transferred information or *transinformation* ..of the channel. Note that, based on the fundamental equation (3-116), the right side of Eq. (3-130) is a nonnegative number. Hence, the average mutual information is also nonnegative, while the individual mutual-information quantities may become negative for some symbol pairs. For a noise-free channel,

$$I(X;Y) = H(X) = H(Y)$$