PROBABILITY AND INFORMATION THEORY, WITH APPLICATIONS TO RADAR

By

P. M. WOODWARD, B.A.

Principal Scientific Officer, Telecommunications Research Establishment, Ministry of Supply

LONDON

PERGAMON PRESS LTD

1953

EDITOR'S PREFACE

THE aim of these monographs is to report upon research carried out in electronics and applied physics. Work in these fields continues to expand rapidly, and it is recognised that the collation and dissemination of information in a usable form is of the greatest importance to all those actively engaged in them. The monographs will be written by specialists in their own subjects, and the time required for publication will be kept to a minimum in order that these accounts of new work may be made quickly and widely available.

Wherever it is practical the monographs will be kept short in length to enable all those interested in electronics to find the essentials necessary for their work in a condensed and concentrated form.

D. W. FRY

AUTHOR'S PREFACE

THE first two chapters of this short monograph are concerned with established mathematical techniques rather than with fresh ideas. They provide the code in which so much of the mathematical theory of electronics and radar is nowadays expressed. Information theory is the latest extension of this code, and I hope that it will not be considered improper that I have tried in Chapter 3 to summarise so much of C. E. SHANNON's original work, which already exists in book-form (*The Mathematical Theory of Communication*, by CLAUDE SHANNON and WARREN WEAVER). The account which is given in Chapter 3 may perhaps spur the reader who has not studied the original literature into doing so.

Chapters 4 and 5 deal with some of the fascinating problems, which have been discussed so often in recent years, of detecting signals in noise. The present approach was suggested to me by SHANNON's work on communication theory and is based on inverse probability; it is my opinion that of all statistical methods, this one comes closest to expressing intuitive notions in the precise language of mathematics. Chapters 5, 6 and 7 are devoted to radar, which is simple enough (ideally) to lend itself to fairly exact mathematical treatment along the lines suggested in the previous chapters. This material is based on papers which have appeared in technical journals, Chapter 6 in particular being a revised account of work originally carried out at T.R.E. in partnership with I.L.Davies. It was this work which led to the present monograph, but it is hoped that the first four chapters—originally conceived as an introduction to the special problems of radar—may find an independent usefulness for the reader whose interests are not so narrowly confined.

I have to thank the Chief Scientist, Ministry of Supply, for permission to publish this book.

P. M. W.

Malvern
March, 1953.

CONTENTS

1

AN INTRODUCTION TO PROBABILITY THEORY

1.1 THE RULES OF PROBABILITY

If n possibilities are equally likely and exactly m of them have some attribute A, we say that the probability of A is m/n. Strictly, this is not a definition of probability because it assumes that the notion of equally likely possibilities is understood in the first place. From a purely mathematical point of view, however, no definition is required. All we need is a set of rules for adding and multiplying probabilities, which are then taken as the basic postulates of the theory. But the study of probability is made easier and the rules become intuitive rather than arbitary when from the beginning there is an obvious practical interpretation, and this the opening remark supplies.

Since probability is a fraction of equally likely possibilities, it is often helpful to set these out in tabular form. Thus

$$A \quad A \quad A \quad B \quad B \quad B \quad B \quad C \tag{1}$$

signifies that $P(A)$, the probability of A, is $\frac{3}{8}$ and so on. It is immediately evident that the probability of A *or* B is $P(A) + P(B)$. This is the *sum rule* and it applies only when A and B cannot simultaneously be true, in other words, when they are mutually exclusive. When all mutually exclusive attributes have been taken into account, their probabilities will naturally add up to unity.

It frequently happens that two sets of attributes, each mutually exclusive among themselves, have to be considered together. Suppose, for instance, that we have eight pencils, three red (A), four black (B) and one blue (C). The scheme (1) represents the equally likely possibilities when one pencil is selected randomly. But the same pencils may also be hard (J) and soft (K) as follows,

$$\begin{matrix} A & A & A & B & B & B & B & C \\ J & J & K & J & J & J & K & K \end{matrix} \tag{2}$$

This scheme signifies that the probability of J is $\frac{5}{8}$ and of K is $\frac{3}{8}$. Suppose now that a chosen pencil is examined for colour only, and found to be A. The information rules out B and C from (2), so the probability of J immediately changes, and becomes $\frac{2}{3}$. It is, therefore, important to state, in relation to any probability, what relevant facts are already known. A brief notation is to write the unconditional probability of J as $P(J)$ but to distinguish the probability when A is given by putting A as a suffix. This brings us to the *product rule* which gives the joint probability of a pair of attributes. The probability of X *and* Y is

$$P(X, Y) = P(X)P_X(Y) = P(Y)P_Y(X) \tag{3}$$

For example, if X is A and Y is J in (2), we have the three equivalent expressions

$$P(A, J) = \tfrac{1}{4}$$
$$P(A)P_A(J) = \tfrac{3}{8} \cdot \tfrac{2}{3}$$
$$P(J)P_J(A) = \tfrac{5}{8} \cdot \tfrac{2}{5}$$

If we have $P(X) = P_Y(X)$, it follows from equation (3) that we also have $P(Y) = P_X(Y)$. Thus knowledge of the one does not affect the probability of the other, and we can say that X and Y are statistically independent. Only then may we write the product rule in the simplified form

$$P(X, Y) = P(X)P(Y) \tag{4}$$

though it is this form which is usually remembered.

The sum and product rules are the main axiomatic foundation upon which the theory of probability rests.

1.2 BERNOULLI'S THEOREM

Of all theorems in probability, Bernoulli's is the one which gives the clearest insight into the behaviour of chance quantities. Suppose that some event is known to have a probability p of occurring whenever a "trial" is made. "Event" and "no event," which we shall denote symbolically by 1 and 0, are the two mutually exclusive attributes. What can we say about the number of events which will occur when n independent trials are made? Intuitively, we should, of course, expect about np events; Bernoulli's theorem confirms this and makes it more precise.

The first step is to consider any arbitrary sequence of results such as 01101. The probability of this particular sequence occurring is given by the product rule for independent attributes, and is $(1-p)pp(1-p)p$. There are

$$^5C_3 = \frac{5 \cdot 4 \cdot 3}{1 \cdot 2 \cdot 3}$$

ways of obtaining three events in five trials when all the different orders are counted, and all have the same probability. The total

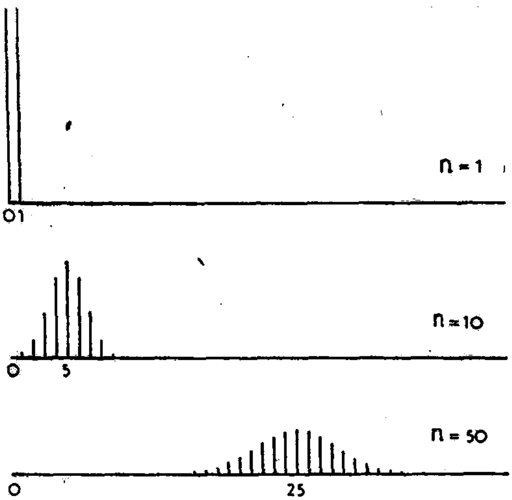

Fig. 1. Symmetrical binomial distributions (Ordinates are probabilities, abscissae number of events, and n the number of trials, each yielding an event with probability 0·5)

probability of obtaining three events in five trials, by the sum rule, is therefore

$$^5C_3 p^3 (1-p)^2$$

The general result, BERNOULLI's theorem, is that the probability of r events in n trials is

$$P_n(r) = {}^nC_r p^r (1-p)^{n-r} \qquad (5)$$

The total probability of obtaining any number of events from 0 to n is thus

$$(1-p)^n + \ldots + {}^nC_r p^r (1-p)^{n-r} + \ldots + p^n \qquad (6)$$

which is equal, as it obviously should be, to unity. It is, in fact, the expansion of

$$[p + (1 - p)]^n \tag{7}$$

by the Binomial theorem.

BERNOULLI's probabilities, more usually called the *binomial distribution*, are illustrated in fig. 1 for $p = 0{\cdot}5$ and $n = 1, 10$ and 50. The probabilities are greatest in the neighbourhood of np events, but the probability of obtaining exactly np events maybe very small. The interesting thing is what happens when n becomes larger and larger, p being held fixed. The peak at np moves to the right in direct proportion to n, the hump gets broader, and the probability of exactly np events must therefore get smaller and smaller, since the sum of all the terms is invariably unity. But the hump does not broaden in direct proportion to the number of trials: its width increases only as the square root of n. Consequently the ratio of events to trials becomes more and more closely equal to p as n is increased. In precise terms, it can be shown that the probability of the number of events lying within the range

$$n(p \pm \varepsilon) \tag{8}$$

tends to unity as n increases, however small ε may be. This is a most important fact: it lies at the root of statistics and of information theory. It means that we can pretend that n trials will give np events with an arbitrarily small fractional error provided n is sufficiently large; this is what is generally meant by the *law of averages*.

1.3 MOMENTS AND GENERATING FUNCTIONS

In most applications of probability theory to physical science, and to electronics in particular, we have to deal with problems in which unit probability is distributed over a set of *quantitative* attributes. The significant feature of quantities—as opposed to qualities—is that they can be ordered, and there is a "distance" between any two of them. The attributes can meaningfully be represented as points on a line, as in fig. 1, or more generally in an attribute-space of any number of dimensions. This gives rise to a geometry of probability distributions in which moments play an important part.

If $P(r)$ is any distribution of probability at discrete points along a line, the nth moment is defined as

$$M_n = \sum_r r^n P(r) \tag{9}$$

The moment of order zero is, of course, unity. The first moment is the centroid of the points r, weighted by the $P(r)$. This is the average value of r which would be obtained when a large number of independent determinations had been made. For if N determinations are made, and N is large enough, r will turn up roughly $NP(r)$ times. The average of the results will therefore be

$$M_1 = \frac{NP(0) \,.\, 0 + NP(1) \,.\, 1 + NP(2) \,.\, 2 + \ldots}{N} \tag{10}$$

In a similar way M_2 is the average value of r^2 (analogous to moment of inertia) and so on. This justifies the more usual notation

$$M_n = \overline{r^n} \tag{11}$$

It will be obvious, more generally, that the average value (or "expectation") of any function $f(r)$ is given by

$$\overline{f(r)} = \sum_r f(r)P(r) \tag{12}$$

The geometrical significance of the second moment, for which $f(r) = r^2$, lies in its relation to the spread of the probabilities about $\bar{r}$. Spread is most naturally measured by the mean squared deviation of r from $\bar{r}$; thus

$$\sigma^2 = \sum_r (r - \bar{r})^2 P(r) \tag{13}$$

$$= \Sigma r^2 P(r) - 2\bar{r}\Sigma r P(r) + (\bar{r})^2 \Sigma P(r)$$

$$= \overline{r^2} - (\bar{r})^2 \tag{14}$$

The quantity σ^2, the second moment minus the squared first moment, is known as the *variance* of the distribution $P(r)$. Its square root, σ, is the *standard deviation*, and gives the width of the hump in distributions such as the binomial. This, however, is a somewhat incautious statement, because the value of σ can depend markedly on the asymptotic behaviour of the very small probabilities a long way from the centroid. Some humped distributions yield an infinite value of σ, but this does not occur very often since most of the probability distributions which arise in physics fall off at least as rapidly as an exponential in the tails. The tails of the distribution then make a very small contribution to the sum (13) and σ is a reasonable measure of spread. If a distribution has more than one hump or is not of any simple type, the mean and standard deviation may not have any simple graphical interpretation, but they still remain useful mathematical parameters.

Although the moments of a distribution may be evaluated directly from equation (9), there is an alternative method of considerable interest. First we form what is known as the *generating function* of $P(r)$, defined by

$$g(x) = \sum_r x^r P(r) \tag{15}$$

Differentiation with respect to x gives

$$g'(x) = \Sigma r x^{r-1} P(r)$$
$$g''(x) = \Sigma r(r-1) x^{r-2} P(r)$$

and so on. If we now put $x = 1$ in $g'(x)$, the first moment is obtained, and in a similar way the second moment may be obtained from $g''(x)$. Thus

$$r = g'(1) \tag{16}$$
$$\overline{r^2} = g''(1) + g'(1) \tag{17}$$
$$\sigma^2 = g''(1) + g'(1) - \{g'(1)\}^2 \tag{18}$$

The method is particularly effective when applied to the binomial distribution $P_n(r)$. We have

$$g(x) = (px + 1 - p)^n$$
$$g'(1) = np$$
$$g''(1) = n(n-1)p^2$$

and hence

$$\bar{r} = np \tag{19}$$
$$\overline{r^2} = n(n-1)p^2 + np \tag{20}$$
$$\sigma^2 = np(1-p) \tag{21}$$

The values of $\bar{r}$ and σ^2 justify, in a rough and ready way, the assertions made in section 1.2. The fact that σ is proportional to the square root of n does not of itself prove the remark at (8), but it suggests it. A complete proof of that result presents no difficulty of principle.

Moments have a specially simple interpretation as applied to the magnitude of an electric current. Consider the idealized current waveform illustrated in fig. 2. The magnitude in each time-cell is supposed to have been selected at random, each one independently. The only magnitudes allowed are integral multiples of unit current, and the probability of a magnitude or r units occurring in any given cell is $P(r)$, say. We then have

$$\bar{r} = \text{d.c. component} \tag{22}$$

$$\overline{r^2} = \text{mean power} \tag{23}$$

Here and elsewhere, the power of a current (or voltage) is to be understood as the power which would be dissipated in a unit resistance. The instantaneous power of any waveform then becomes conveniently the square of the current or voltage.

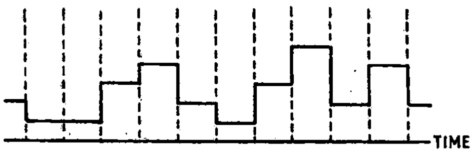

Fig. 2. Idealised noise waveform

The equation (14) for variance in terms of moments may now be expressed as follows

$$\sigma^2 = \text{mean fluctuation power} = (\text{total mean power}) - (\text{d.c. power}) \tag{24}$$

When the random fluctuation is unwanted, it is "noise," and the variance of its magnitude distribution is the mean noise power. Random fluctuations are not always noise; indeed, a communication signal may fluctuate in what appears to be a random manner, and the main feature of communication theory, as we shall see later, is that signals and noise can both be treated as statistical phenomena.

1.4 CONVOLUTION

Suppose that we have to deal with a pair of independent random quantities: the first one is always a whole number r and has the probability distribution $P(r)$, the second is a whole number s with distribution $Q(s)$. A most important problem is to find the probability distribution for $r + s$.

The solution is quite straightforward. Let the sum of r and s be denoted by u, and consider a fixed value of u. If the first quantity is r, the second must be $u - r$, and the probability of obtaining these two particular values is given by the product rule as

$$P(r)Q(u - r) \tag{25}$$

The probability of obtaining any given value of u is the sum of the probabilities for all the different ways in which it can be made up, i.e. the sum of all products like (25) with r varying. Thus the required distribution is

$$R(u) = \sum_r P(r)Q(u - r) \tag{26}$$

It should be pointed out that whilst the symbol P is often used in a purely operational sense, meaning "the probability of . . . ," the symbols P, Q and R in (26) stand for definite mathematical functions representing the probabilities of r, s and u respectively. Mathematically, equation (26) is a way of putting two functions together and forming a kind of resultant. In fact R is sometimes described as the "resultant" of P and Q, but the term *convolution* is more usual and less ambiguous. The following notation will be used:

$$R = P{\star}Q \tag{27}$$

It is very easy to show that the arguments r and $u - r$ in equation (26) may be interchanged, and hence

$$P{\star}Q = Q{\star}P \tag{28}$$

This commutative property is of course quite obvious in terms of the original problem. It is also worth remarking that if P, Q and S are three functions, we have

$$P{\star}(Q{\star}S) = (P{\star}Q){\star}S \tag{29}$$

and a unique significance therefore attaches to $P{\star}Q{\star}S$ without brackets. It would be the distribution function for the sum of three quantities, which makes the associative property obvious.

Let us now examine the convolution formula in more detail. Suppose that the random quantity r can only assume the values 0, 1, 2 or 3 and that s can only be 0, 1 or 2. Then we have

$$\begin{aligned}
R(0) &= P(0)Q(0) \\
R(1) &= P(0)Q(1) + P(1)Q(0) \\
R(2) &= P(0)Q(2) + P(1)Q(1) + P(2)Q(0) \\
R(3) &= \phantom{P(0)Q(2) + {}} P(1)Q(2) + P(2)Q(1) + P(3)Q(0) \\
R(4) &= \phantom{P(0)Q(2) + P(1)Q(2) + {}} P(2)Q(2) + P(3)Q(1) \\
R(5) &= \phantom{P(0)Q(2) + P(1)Q(2) + P(2)Q(2) + {}} P(3)Q(2)
\end{aligned}$$

This can be set out in the form of a long multiplication,

$Q(0)$	$Q(1)$	$Q(2)$			
$P(0)$	$P(1)$	$P(2)$	$P(3)$		
$P(0)Q(0)$	$P(0)Q(1)$	$P(0)Q(2)$			
	$P(1)Q(0)$	$P(1)Q(1)$	$P(1)Q(2)$		
		$P(2)Q(0)$	$P(2)Q(1)$	$P(2)Q(2)$	
			$P(3)Q(0)$	$P(3)Q(1)$	$P(3)Q(2)$
$R(0)$	$R(1)$	$R(2)$	$R(3)$	$R(4)$	$R(5)$

(30)

It should be noticed that the "multiplication" is a perfectly genuine algebraic product of the generating functions of P and Q. Thus we have the following important equivalents:

$$\left.\begin{array}{l}\text{Sum of random quantities}\\ \text{Convolution of probability distributions}\\ \text{Product of generating functions}\end{array}\right\} \qquad (31)$$

A further property is that the mean and variance of each quantity is additive under convolution. Thus if $g(x)$ is the generating function of $P{\star}Q$, we have

$$g(x) = \sum_r x^r P(r) \sum_s x^s Q(s)$$

Differentiating,

$$g'(x) = \sum_r r x^{r-1} P(r) \sum_s x^s Q(s) + \sum_r x^r P(r) \sum_s s x^{s-1} Q(s)$$

By equation (16), the mean value of u is obtained by putting $x = 1$, and hence

$$\bar{u} = \bar{r} + \bar{s} \qquad (32)$$

which is the first result. By further differentiation one can show, using equation (18), that

$$\sigma_u^2 = \sigma_r^2 + \sigma_s^2 \qquad (33)$$

where σ_u^2 is the variance of u, etc. These two results are of extreme usefulness.

The binomial distribution illustrates all the above very satisfactorily. In section 1.2 we considered a trial of which the possible outcomes were 1 and 0, with probabilities p and $1 - p$ respectively. We may therefore write

$$\left.\begin{aligned} P(0) &= 1 - p \\ P(1) &= p \end{aligned}\right\} \tag{34}$$

and the generating function of this simple distribution is

$$1 - p + px \tag{35}$$

The number of 1's which will be obtained when n independent trials are made, is the sum of the results of each trial. Each trial has the same distribution (34) and the same generating function (35). Thus the distribution for the number of 1's in n trials has, by rule (31), the generating function

$$(1 - p + px)^n \tag{36}$$

which gives immediately the binomial distribution of order n. Its mean and variance, np and $np(1 - p)$ evaluated in section 1.3, are each proportional to n, confirming equations (32) and (33).

1.5 EVENTS AT RANDOM IN TIME

Events at random in time frequently occur in physical problems, and although not strictly necessary in the present monograph the simple theory of them can hardly be passed over in this introduction.

Let time be divided into small intervals each of duration δt, and let the probability of an event occurring within any interval be p, independently for each interval. For example, if $p = 0{\cdot}1$, a sequence such as the following might be obtained

000010000100000001000000000000000000000
000000010110100000000000111000100100000000

The number of events r in unit time obeys a binomial distribution $P_n(r)$, where n is the number of trials. Suppose now that the intervals δt are made shorter and the probability p per interval proportionately reduced such that the mean number of events in unit time is always fixed and equal, say, to λ. In the limit, we obtain events truly at random in time. Thus, in unit time, we have

$$\bar{r} = np = \lambda \tag{37}$$

whilst

$$n \to \infty, p \to 0 \tag{38}$$

The generating function of $P_n(r)$ becomes, from (36) and (37)

$$g(x) = \lim_{n\to\infty}\left\{1 + \frac{\lambda}{n}(x-1)\right\}^n$$
$$= e^{\lambda(x-1)} \tag{39}$$

The coefficient of x^r in the expansion of $g(x)$ is, by definition, the probability of r events. Thus we obtain

$$P(r) = \frac{\lambda^r}{r!}e^{-\lambda} \tag{40}$$

which is the POISSON distribution. The mean and variance are both equal to λ, from (19), (21) and (38).

A further distribution of interest is the one for the time which elapses between successive events. Strangely, as it seems at first, this distribution is the same as that which describes the time

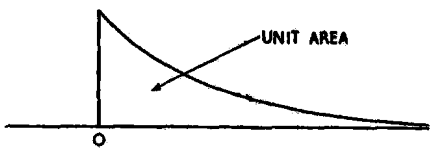

Fig. 3. The exponential distribution

between any instant selected at random and the next event. The probability that any interval in the discrete sequence will be followed by $s - 1$ zeros and then by an event is

$$P(s) = (1 - p)^{s-1}p \tag{41}$$

where, from equation (37), we have

$$p = \lambda\delta t \tag{42}$$

since $n = 1/(\delta t)$. In another form, the probability that the next event occurs within an interval δt a time t later may be denoted by $q(t)\delta t$, where $t = s\delta t$. Then we have

$$q(t)\delta t = \left(1 - \frac{\lambda t}{s}\right)^{s-1}\lambda\delta t$$

and going to the limit $s \to \infty$, we have the *exponential distribution* (fig. 3)

$$q(t)dt = \lambda e^{-\lambda t}dt, \qquad t > 0 \tag{43}$$