1

PROBABILITY

Introduction

The object of this book is to provide an introduction to the study of random phenomena. Most phenomena studied in science are not looked at from a random viewpoint but from a deterministic one, and may often be put into the simple form of cause and effect: if A then B. When randomness is present A may sometimes cause B, but sometimes C. Nevertheless, random phenomena, like deterministic ones, are capable of mathematical description; and in this book we show how this can be done using probability ideas. There are two aspects to the study. In the first half of the book, chapters 1–4, we deal with several simple random phenomena combining together to produce a more complicated random situation: the argument is purely deductive. In the second half, chapters 5–8, inductive problems are considered. If A_i always causes B_i, and no two of the B_i are the same, then it is a simple inductive process to infer that if B_i is observed then A_i must have been the cause. But if A_i can sometimes cause B_i and sometimes B_j, then the cause of B_j may be either A_i or A_j and the induction is not so complete. Nevertheless, again using probability ideas, inductive statements can be made precise. We shall often use the term 'statistics' to cover this latter study and reserve the term 'probability' for the purely deductive part.

Examples of deterministic statements of cause and effect are: if, under specified conditions, an electric current is passed through water, hydrogen and oxygen will be emitted; if the temperature of water is lowered sufficiently it will freeze; if a penny is released from the hand it will fall to the floor, and it can be calculated how fast it will fall. In contrast there are several situations in which it does not seem possible to make such statements of cause and effect, or if they were made they would be too complex to be useful. A simple example is provided by the release of a penny. If the penny is given a spinning motion it is still true that the penny will fall but it is not possible to say whether the effect will be that it will come to rest with the head or with the tail uppermost. Even if the initial conditions, or causes, were described with great care so that it would, in theory, be possible to predict the exposed face, it would not be useful to do so. Other, more practical, examples are commonplace. It is not possible to say whether or not an Englishman aged 40 will die within 10 years. It is not possible to say what the weather will be on a specified day a year hence. It is not possible to say what the hair colour of the child of brunette parents will be. It is not possible to say whether a given atom of uranium will disintegrate within 10 years. Here are several initial conditions but we cannot say what will be the corresponding effects. Although statements of the form, if A then B, cannot be made, important statements of a different kind can be made: witness, in the four examples quoted, the use of actuarial life tables, statements about climate, the science of genetics and the theory of radioactive decay.

Granted that it is possible to make precise statements about the toss of a penny (the way this can be done is described in the following sections) it is possible to deduce other results. For example, it can be calculated how long one can expect to go on tossing until the numbers of heads and tails are equal (§4.4); or how many heads one can expect to get in ten tosses (§2.1). This is a matter of mathematical deduction like deducing that the sum of the angles of a triangle is two right angles from the postulated properties of lines and points. From the random way in which telephone calls are made we can deduce the demand on an exchange to which a group of persons are connected (§§4.2, 4.3). We can answer questions such as what will be the effect of providing more telephone operators? These are deductive probability problems. Consider next a problem which is, in a sense, converse to the problem of the number of heads in ten tosses of a coin. If a penny has been observed to give eight heads in ten tosses is it likely to be a fair one obtained from a reputable mint? By fair we mean, to anticipate terms to be introduced later, that it is 'equally likely' to come down heads or tails in a single toss. From the fairness of a single toss the number of heads to be expected in ten tosses can be *deduced*, as just mentioned. To go in the reverse direction, from the results of the tosses to the fairness, means an *inductive* process, and is a problem typical of statistics. It is discussed in §7.2. Examples of more practical statistical problems are, to infer from agricultural experiments (which are subject to substantial random variation) the relative merits of different strains of wheat (§§6.4, 6.5); to compare the potencies of different drugs in an experiment on animals; to assess the accuracy of determination of a physical or chemical constant from an experiment which, like most experiments, must contain some randomness; and to detect and measure linkage between two genes.

1.1. The concept of a frequency limit

Consider a situation in which A can either cause an event B to occur or not, so that the situation is not deterministic. We say A either produces B or not-B, which we denote by $\bar{B}$. In many such situations it is possible to repeat A and observe, on each repetition, B or $\bar{B}$. It is an empirical fact that often, as the number, n, of repetitions increases, the ratio of the number, m, of times B occurs to the total number of repetitions becomes stable and appears to tend to a limit. Each repetition is termed a *trial*, the occurrence of the event B is a *success* (and of $\bar{B}$, a *failure*) and the ratio m/n the *success* (or *frequency*) *ratio*. The empirical observation can be expressed by saying that '$\lim_{n\to\infty}$' m/n exists, where the limit symbol has been placed in quotation marks to indicate that the notion is not the same as the ordinary mathematical limit. The limiting number is the empirical value of the *probability of B given A*, which is written $p(B|A)$. Any probability is a function of two arguments which are separated by a vertical line; the first is the event being considered, the second describes the conditions under which it is being considered and is called the *conditioning event*, or simply the *condition*: here, the event 'the occurrence of B' is being considered under condition A. In the next section axioms for probability suggested by this empirical fact will be given.

Empirical results

The simplest examples of repetitions of trials are provided by games of chance. If A denotes the toss of a penny and B denotes the fall of it with head uppermost, then repetitions are possible and the empirical fact observed. The limit is the probability of head on the toss of a coin. Kerrich (1946) carried out numerous experiments in order to demonstrate the stability of the success ratio: for example, he spun a coin 10,000 times and demonstrated that the ratio for 'heads' kept very near to 1/2.

If A denotes the roll of a die and B the occurrence of some number, or set of numbers, then again the empirical limit can be observed. Weldon rolled some dice and exhibited the stability of the success ratio for the event B, 'a 5 or 6'. The limit was not 1/3 as one might expect for a 'fair' die but somewhat more. It is relevant to notice that there is nothing in the above statement about the existence of the limit to say what the limit is. For a newly minted coin in our first example the limit is probably 1/2, but for a badly bent one it might well, as with Weldon's dice, be different from the ideal value.

The ultimate justification for probability does not lie in such experiments but rather in the practical success with which the theory has been applied. The prosperity of insurance houses is a witness to the value of statements about the probability of death. The science of genetics is based on the same sort of head or tail phenomenon as coin-tossing: will the gene transferred from parent to offspring be B or $b\,(=\bar{B})$? A breeding programme will provide the repetitions in which the stability can be observed, but this is secondary to the success of deductions from the theory. Games of chance provide the oldest example. The Chevalier de Méré played so many games that he was able, on empirical evidence alone, to detect that a certain limiting success ratio was less than 1/2: the mathematician Pascal showed that it was 0·491 by suitable application of probability theory. In radioactive studies the 'half-life' term is another way of expressing the limit: in repeated observations on different atoms, a success, namely decay, will have been observed in about 1/2 of them after the lapse of time equal to the 'half-life'.

The conditioning event

It is important to notice that the conditioning event A is just as relevant to the probability as is the event B. Consider tossing a coin where the event B, of success, is a head. Then, if the coin is a newly minted one, the probability will be about 1/2, but if the coin is badly bent, or if a piece of chewing-gum is stuck on to one face, the probability of the same event may be very far from 1/2. It is easy to produce paradoxes by failing to mention the conditioning event. Of course, in a series of trials it is impossible to keep the conditioning event completely constant. At least the time the trials are carried out will be different and typically the coin or die will show wear. But such difficulties of precision always arise in discussing the relationship between theory and practice: it is useful to think of an object as having a fixed weight, for example in use on a balance, but this is not precisely true. Conditions for the limit to exist within the theory will be given in §3.6 (theorems 2 and 3).

The form of 'lim'

The limit here is not a mathematical limit. That is to say, given any small positive number ϵ, it is not possible to find a value N such that $|m/n-p| < \epsilon$ for all $n > N$, where $p = \text{`lim'}_{n\to\infty} m/n$, as would be required of a mathematical limit. For there is nothing impossible in m/n differing from p by as much as 2ϵ, it is merely rather unlikely. And the word unlikely involves probability ideas so that the attempt at a definition of 'limit' using the mathematical limit becomes circular. The axiomatic approach (§1.2) avoids the difficulty, and the empirical observation will not be used to define probability, but only to suggest the axioms.

Examples

The probability statements that can be made in the examples used in the introduction are the following. The probability that an Englishman aged 40 will die within 10 years is 0·05: the event B is death within 10 years, A is the condition 'an Englishman aged 40'. The probability of rain at a specified place on a given date can be found from the rainfall statistics. The probability that the offspring of heterozygous parents will exhibit the recessive phenotype is $\frac{1}{4}$. The probability of decay of a uranium atom within $4{\cdot}49 \times 10^9$ years is $\frac{1}{2}$. All these results are based on the observation of repeated trials, supported, in the genetic case particularly, by additional indirect evidence.

1.2. The axioms of probability

The notion of an event is first formalized, and then the notion of the probability of an event. We consider a *sample space*, $\mathbf{A}$, consisting of points, a, called *elementary events*. An *event* is a collection or *set* of elementary events and is denoted by a capital letter A, B, C, ..., with suffixes A_1, A_2, ..., where necessary. If a belongs to A we write $a \in A$. Selection of a particular a is referred to by saying 'a has occurred'. If $a \in A$ and a has occurred we say A has occurred. If A and B are two events, the set of a such that both $a \in A$ and $a \in B$ is denoted by AB. If AB has occurred then both A and B have occurred, and conversely. If $\{A_n\}$ is a sequence of events, the set of a which belong to at least one A_n is denoted by $\sum_n A_n$. If $\sum_n A_n$ has occurred then at least one A_n has occurred, and conversely. The members of the sequence are *exclusive given* C, if whenever C has occurred no two of them can occur together, that is if $A_m A_n C$ is the empty set whenever $m \neq n$. If the conditioning event C is $\mathbf{A}$, the sample space, then, in this last definition, and similar ones, the words 'given $\mathbf{A}$' are omitted. If $\sum_n A_n = \mathbf{A}$, that is, if every a belongs to at least one A_n, then $\{A_n\}$ is said to be *exhaustive*.

For certain pairs of events, A and B, a real number $p(A|B)$ is defined and called the *probability* of A given B. These numbers satisfy the following axioms:

Axiom 1. $0 \leqslant p(A|B) \leqslant 1$ and $p(A|A) = 1$.

Axiom 2. If the events in $\{A_n\}$ are exclusive given B then

$$p(\sum_n A_n | B) = \sum_n p(A_n|B).$$

Axiom 3 $p(C|AB)\, p(A|B) = p(AC|B)$.

The role of the axioms

The principle behind the construction of any axiom system is that a mathematical representation should be made of certain aspects of the real world. The elements in the mathematics are not parts of the real world but only representations of them. The elements are given properties (called axioms) supposed to reflect the behaviour of the corresponding parts of the real world. These properties can then be used in conjunction with the rules of mathematical logic to deduce other properties (within the mathematical system) which may be compared with the real world. If the axiomatization has been successful the comparison will lead to fruitful new ideas. The classical example is Euclid's geometry with axioms about lines and points, etc., such as 'through two points passes a unique line' and deductions like 'the sum of the angles of a triangle is two right angles'. Let us consider the axiom system described above in relation to the real phenomenon of tossing a penny.

The sample space

The event of a penny being tossed is more complicated than a mere occurrence of heads or tails: we could also consider its position of fall, the time of fall, etc., indeed countless other facets of the toss. Any particular toss may be represented by a point a and all possible tosses form the sample space. The collection of those a which result in heads is the event of heads having occurred. Any elementary event then will contain details of what penny was tossed, when and where it was tossed, etc. It is not necessary to be explicit in saying what is or is not contained in the description of an elementary event: it can be thought of quite abstractly as the toss being considered. It is often helpful to represent each elementary event by a point on the paper, and an event by a region of the paper containing the points representing the elementary events contained in the event. If A and B are two events so represented then the event $A+B$ is represented by the region which consists of the sum of the two regions for A and B. The event AB is represented by the region common to that for A and that for B. Thus in fig. 1.2.1 the event

A is represented by the horizontal rectangle: the event B by the vertical rectangle. $A+B$ is represented by the T-shaped figure and AB by the shaded square. Such a representation is called a *Venn diagram*. It will be found useful in understanding the proofs of the theorems in the next section.

Empirical justification for the axioms

Consider an event consisting of all tosses with a given coin under a standard set of conditions. Denote this event by B and the event of heads by A. Then $p(A|B)$ is the mathematical representation of the frequency limit discussed in §1.1. To see

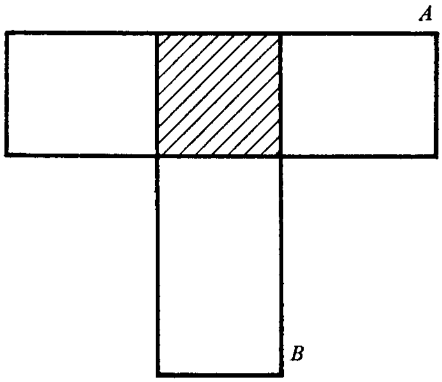

Fig. 1.2.1. Venn diagram for two events, A and B.

this, consider each axiom in turn. The frequency limit obviously lies between 0 and 1 and if we confine attention to tosses resulting in heads, so that the conditioning event is 'heads', then heads will always result and the limit necessarily be one, so that in any representation $p(A|A)$ must be one. This explains the first axiom. To appreciate the second consider N different events $A_1, A_2, \ldots, A_N$ concerning the outcome of the toss which are such that, given B, no two can occur together (they are exclusive, given B). For example, let A_i be the event that in the final rest position of the penny the acute angle between a fixed line on the head of the penny and a fixed line on the table lies between $(i-1)\pi/2M$ and $i\pi/2M$, the former limit being included and the latter excluded and M exceeding N. Then if, in n tosses,

A_i occurs m_i times, the event $\sum_i A_i$ (at least one A_i occurs) occurs $\sum_i m_i$ times. (This would not necessarily be true if the A_i were not exclusive.) The success ratio for $\sum_i A_i$ is $\sum_i m_i/n$, the sum of the success ratios for the individual events. Since the success ratios have this property it is reasonable to assume the same for their limits. Hence the second axiom. It is mathematically convenient to suppose this property holds for a countably infinite number of events† as well as for a finite number; the right-hand side will then be an infinite series, converging to the value on the left.

The first two axioms apply with a fixed conditioning event B. This is not so with the third axiom where two conditioning events B and AB both occur. To interpret this axiom let A and B be as before and let C be the event A_1, say, just referred to. Consider n tosses with the given penny under standard conditions, that is n occurrences of event B. Suppose there are m heads, A occurs m times. Suppose that amongst those m occasions C also occurs on r of them; so that r is the number of times the event AC occurs. Then trivially we have

$$\left(\frac{r}{m}\right)\left(\frac{m}{n}\right) = \left(\frac{r}{n}\right).$$

Now let $n \to \infty$, and hence $m \to \infty$ unless $p(A \mid B) = 0$. Then

$$\text{`}\lim_{m\to\infty}\text{'}\,\frac{r}{m} = p(C \mid AB),$$

since the conditioning event is AB which occurs m times, and the observed event is C which occurs on r of these m occasions;

$$\text{`}\lim_{n\to\infty}\text{'}\,\frac{m}{n} = p(A \mid B) \quad \text{and} \quad \text{`}\lim_{n\to\infty}\text{'}\,\frac{r}{n} = p(AC \mid B) \quad \text{similarly.}$$

Consequently the third axiom is a reasonable limiting interpretation of the trivial result. If $p(A \mid B) = 0$, then necessarily $p(AC \mid B) = 0$ since AC is of rarer occurrence than A, and the result persists in that case. This completes the justification for

† The number of events (or other concepts) is countably infinite if they can be put into one-to-one correspondence with the integers 1, 2, ...

the axioms in the coin-tossing case: the reader might like to carry through one of the other examples similarly.

The reader may wonder why the complicated sample space of elementary events was introduced at all. Certainly, in simple problems like heads or tails in tossing pennies, it is not essential to have more than two distinct elementary events, 'heads' and 'tails' and it is often enough to use this sample space. But it is an advantage to consider the full sample space because then any event (such as C above) can also be discussed without changing the sample space. Had the sample space of only two elements been used, it would have had to have been altered before C could be discussed. Since the elementary events need not be formulated explicitly there is great advantage and little real addition in complexity in enlarging the sample space to its full extent. Furthermore, in defining the probabilities, it is only necessary to consider the events, and not the elementary events, so that the introduction of complicated elementary events does not increase the difficulty of defining the probabilities. Thus, in coin tossing, however involved the description of the individual tosses be, only the events of heads and tails need be considered.

Fixed conditions

It often happens throughout a probability calculation that one event B always occurs as part or the whole of the conditioning event: for example, the event B just defined as tosses with a single coin under standard conditions. B is often the whole sample space, $\mathbf{A}$, or may be taken to be that (see the comment on theorem (1.4.1) below). It then economizes on notation to write $p(A)$ for $p(A|\mathbf{A})$ and $p(A \mid C)$ for $p(A \mid \mathbf{A}C)$, omitting reference to $B = \mathbf{A}$. Indeed most writers on probability define probability as $p(A)$, a function of a single event: but this can be misleading and it pays to remember the conditioning event. To omit it is rather like considering an effect without its cause. (Compare the discussion in §1.1.)

One method of constructing a probability system is as follows. With certain events, A, associate a real number $p(A)$ satisfying

Axiom 1*a*. $0 \leqslant p(A) \leqslant 1 \quad \text{and} \quad p(\mathbf{A}) = 1.$