1

Introduction to probability theory

Unpredictability and non-determinism are all around us. The future behaviour of any system—from an elementary particle to a complex organism—may follow a number of possible paths. Some of these paths may be more likely than others, but none is absolutely certain. Such unpredictable behaviour, and the phenomena that cause it, are usually described as 'random'. Whether randomness is in the nature of reality, or is the result of imperfect knowledge, is a philosophical question which need not concern us here. More important is to learn how to deal with randomness, how to quantify it and take it into account, so as to be able to plan and make rational choices in the face of uncertainty.

The theory of probabilities was developed with this object in view. Its domain of applications, which was originally confined mainly to various games of chance, now extends over most scientific and engineering disciplines.

This chapter is intended as a self-contained introduction; it describes all the concepts and results of probability theory that will be used in the rest of the book. Examples and exercises are included. However, it is impossible to provide a thorough coverage of a major branch of mathematics in one chapter. The reader is therefore assumed to have encountered at least some of this material already.

1.1 Sample points, events and probabilities

We start by introducing the notion of a 'random experiment'. Any action, or sequence of actions, which may have more than one possible outcome, can be considered as a random experiment. The set of all possible outcomes is usually denoted by Ω and is called the 'sample space' of the experiment. The individual outcomes, or elements of Ω, are called 'sample points'.

The sample space may be a finite or an infinite set. In the latter case, it may be enumerable (i.e. the outcomes can be numbered $0, 1, \ldots$), or it may be non-enumerable (e.g. Ω can be an interval on the real line).

Examples

1. A race takes place between n horses, with the sole object of determining the winner. There are n possible outcomes, so the sample space can be identified with the set $\Omega = \{1, 2, \ldots, n\}$.

2. In the same race, the finishing position of every horse is of interest and is recorded. The possible outcomes are now the $n!$ permutations of the integers $\{1, 2, \ldots, n\}$: $\Omega = \{(1, 2, \ldots, n), \ldots, (n, n-1, \ldots, 1)\}$ (assuming that two horses cannot arrive at the finishing line at exactly the same time).

3. A certain task is carried out by an unreliable machine which may break down before completion. If that happens, the machine is repaired and the task is restarted from the beginning. The experiment ends when the task is successfully completed; the only output produced is the number of times that the task was attempted. The sample space here is the set of all positive integers: $\Omega = \{1, 2, \ldots\}$.

4. For the same machine, the experiment consists of measuring, with infinite accuracy, the period of time between one repair and the next breakdown. The sample points now are the positive real numbers: $\Omega = \{x : x \in \mathcal{R}^+\}$.

* * *

The next important concept is that of an 'event'. Intuitively, we associate the occurrence of an event with certain outcomes of the experiment. For instance, in example 1, the event 'an even-numbered horse wins the race' is associated with the sample points $\{2, 4, 6, \ldots, n - (n \bmod 2)\}$. In example 3, the event 'the task is run unsuccessfully at least k times' is represented by the sample points $\{k+1, k+2, \ldots\}$. In general, an event is defined as a subset of the sample space Ω. Given such an event, A, the two statements 'A occurs' and 'the outcome of the experiment is one of the points in A' have the same meaning.

1.1.1 An algebra of events

The usual operations on sets—complement, union and intersection—have simple interpretations in terms of occurrence of events. If A is an event, then the complement of A with respect to Ω (i.e. those points in Ω which are not in A) occurs when A does not, and vice versa. Clearly, that complement should also be an event. It is denoted by $\overline{A}$, or $\neg A$, or A^c. If A and B are two events, then the union of A and B (the sample points which belong to either A, or B, or both) is an event which occurs when either A occurs, or B occurs, or both. That event is denoted by $A + B$, or $A \cup B$. Similarly, the intersection of A and B (the sample points which belong to both A and B) is an event which occurs when both A and B occur. It is denoted by AB, or $A \cap B$, or A, B.

There is considerable freedom in deciding which subsets of Ω are to be called 'events' and which are not. It is necessary, however, that the definition should be such that the above operations on events can be carried out. More precisely, the set of all events, $\mathcal{A}$, must satisfy the following three axioms.

E1: The entire sample space, Ω, is an event (this event occurs no matter what the outcome of the experiment).
E2: If A is an event, then $\overline{A}$ is also an event.
E3: If $\{A_1, A_2, \ldots\}$ is any countable set of events, then the union $A = \bigcup_{i=1}^{\infty} A_i$ is also an event.

From E1 and E2 it follows that the empty set, $\emptyset = \overline{\Omega}$, is an event (that event can never occur). From E2 and E3 it follows that if $\{B_1, B_2, \ldots\}$ is a countable set of events, then the intersection

$$B = \bigcap_{i=1}^{\infty} B_i = \overline{\bigcup_{i=1}^{\infty} \overline{B_i}}\ , \tag{1.1}$$

is also an event.

The 'countable set' mentioned in E3 may of course be finite:

> If $A_1, A_2, \ldots, A_n$ are events, then the union $A = \bigcup_{i=1}^{n} A_i$ is also an event.

Similarly, (1.1) applies to finite intersections.

In set theory, a family which satisfies E1–E3 is called a σ-algebra (or a σ-field, or a Borel field). Thus the set of all events, $\mathcal{A}$, must be a σ-algebra. At one extreme, $\mathcal{A}$ could consist of Ω and $\emptyset$ only; at the other, $\mathcal{A}$ could contain every subset of Ω.

Two events are said to be 'disjoint' or 'mutually exclusive' if they cannot occur together, i.e. if their intersection is empty. More than two events are disjoint if every pair of events among them are disjoint. A set of events $\{A_1, A_2, \ldots\}$ is said to be 'complete', or to be a 'partition of Ω', if (i) those events are mutually exclusive and (ii) their union is Ω. In other words, no matter what the outcome of the experiment, one and only one of those events occurs.

To illustrate these definitions, consider example 3, where a task is given to an unreliable machine to be carried out. Here we can define $\mathcal{A}$ as the set of all subsets of Ω. Two disjoint events are, for instance, $A = \{1, 2, 3\}$ (the task is completed in no more than three runs) and $B = \{5, 6\}$ (it takes five or six runs). However, if we include, say, event $C = \{6, 7, \ldots\}$ (the task needs at least six runs to complete) then the three events A, B, C are not disjoint because B and C are not. The events A and C, together with $D = \{4, 5\}$, form a partition of Ω.

1.1.2 Probabilities

Having defined the events that may occur as a result of an experiment, it is desirable to measure the relative likelihoods of those occurrences. This is done by assigning to each event, A, a number, called the 'probability' of that event and denoted by $P(A)$. By convention, these numbers are normalized so that the probability of an event which is certain to occur is 1 and the probability of an event which cannot possibly occur is 0. The probabilities of all events are in the (closed) interval $[0, 1]$. Moreover, since the probability is, in some sense, a measure of the event, it should have the additive property of measures: just as the area of the union of non-intersecting regions is equal to the sum of their areas, so the probability of the union of disjoint events is equal to the sum of their probabilities.

Thus, probability is a function, P, defined over the set of all events, whose values are real numbers. That function satisfies the following three axioms.

P1: $0 \leq P(A) \leq 1$ for all $A \in \mathcal{A}$.
P2: $P(\Omega) = 1$.

P3: If $\{A_1, A_2, \ldots\}$ is a countable (finite or infinite) set of *disjoint* events, then $P[\bigcup_{i=1}^{\infty} A_i] = \sum_{i=1}^{\infty} P(A_i)$.

Note that Ω is not necessarily the only event which has a probability of 1. For instance, consider an experiment where a true die is tossed infinitely many times. We shall see later that the probability of the event 'a 6 will appear at least once' is 1. Yet that event is not equal to Ω, because there are outcomes for which it does not occur. In general, if A is an event whose probability is 1, then A is said to occur 'almost certainly'.

An immediate consequence of P2 and P3 is that, if $\{A_1, A_2, \ldots\}$ is a (finite or infinite) partition of Ω, then

$$\sum_{i=1}^{\infty} P(A_i) = 1 . \tag{1.2}$$

In particular, for every event A, $P(\overline{A}) = 1 - P(A)$. Hence, the probability of the empty event is zero: $P(\emptyset) = 1 - P(\Omega) = 0$. Again, it should be pointed out that this is not necessarily the only event with probability 0. In the die tossing experiment mentioned above, the probability of the event '6 never appears' is 0, yet that event may occur.

It is quite easy to construct a probability function when the sample space is countable. Indeed, suppose that the outcomes of the experiment are numbered $\omega_1, \omega_2, \ldots$. Assign to ω_i a non-negative weight, p_i ($i = 1, 2, \ldots$), so that

$$\sum_{i=1}^{\infty} p_i = 1 . \tag{1.3}$$

Then the probability of any event can be defined as the sum of the weights of its constituent sample points. This definition clearly satisfies axioms P1–P3. Consider again example 3: one possibility is to assign to sample point $\{i\}$ weight $1/2^i$. The events mentioned above, $A = \{1, 2, 3\}$, $B = \{5, 6\}$, $C = \{6, 7, \ldots\}$ and $D = \{4, 5\}$ would then have probabilities $P(A) = 7/8$, $P(B) = 3/64$, $P(C) = 1/32$ and $P(D) = 3/32$ respectively. Note that the probabilities of A, C and D (those three events form a partition of Ω) do indeed sum up to 1.

When the sample space is uncountable (like the positive real axis in example 4), it is more difficult to give useful definitions of both events and probabilities. To treat that topic properly would involve a considerable excursion into measure theory, which is outside the scope of this book. Suffice to say that the events are the measurable subsets of Ω

and the probability function is a measure defined over those subsets. We shall assume that such a probability function is given.

If A and B are two arbitrary events, then

$$P(A \cup B) = P(A) + P(\bar{A}B) . \tag{1.4}$$

This is a consequence of the set identity $A \cup B = A \cup (\bar{A}B)$, plus the fact that A and $\bar{A}B$ are disjoint. Also, from $B = (AB) \cup (\bar{A}B)$ it follows that $P(B) = P(AB) + P(\bar{A}B)$. Hence,

$$P(A \cup B) = P(A) + P(B) - P(AB) . \tag{1.5}$$

In general, if $A_1, A_2, \ldots$ are arbitrary events, then

$$P\left[\bigcup_{i=1}^{\infty} A_i\right] \leq \sum_{i=1}^{\infty} P(A_i) . \tag{1.6}$$

The inequality in (1.6) becomes an equality only when $A_1, A_2, \ldots$ are disjoint.

The probability of the intersection of two events is not necessarily equal to the product of their probabilities. If, however, that happens to be true, then the two events are said to be independent of each other. Thus, A and B are independent if

$$P(AB) = P(A)P(B) . \tag{1.7}$$

As an illustration, take example 2, where n horses race and there are $n!$ possible outcomes. Let $\mathcal{A}$ be the set of all subsets of Ω and the probability function be generated by assigning to each of the $n!$ outcomes probability $1/n!$ (i.e. assume that all outcomes are equally likely). Suppose that $n = 3$ and consider the following events:

$A = \{(1,2,3),(1,3,2)\}$ (horse 1 wins);
$B = \{(1,2,3),(2,1,3),(2,3,1)\}$ (horse 2 finishes before horse 3);
$C = \{(1,2,3),(1,3,2),(2,1,3)\}$ (horse 1 finishes before horse 3).

Then events A and B are independent of each other, since

$$P(AB) = P(\{(1,2,3)\}) = 1/6 ,$$

and

$$P(A)P(B) = (2/6)(3/6) = 1/6 .$$

However, events A and C are dependent, because

$$P(AC) = P(\{(1,2,3),(1,3,2)\}) = 2/6 ,$$

while

$$P(A)P(C) = (2/6)(3/6) = 1/6 \, .$$

It is equally easy to verify that events B and C are dependent.

The above definition of independence reflects the intuitive idea that two events are independent of each other if the occurrence of one does not influence the likelihood of the occurrence of the other. That definition is extended recursively to arbitrary finite sets of events as follows: the n events $A_1, A_2, \ldots, A_n$ $(n > 2)$ are said to be 'mutually independent', if both of the following conditions are satisfied:

(i) $P(A_1 A_2 \ldots A_n) = P(A_1)P(A_2) \ldots P(A_n)$.
(ii) Every $n - 1$ events among the $A_1, A_2, \ldots, A_n$ are mutually independent.

It should be emphasized that neither the first nor the second condition by itself is sufficient for mutual independence. In particular, it is possible that independence holds for every pair of events, yet does not hold for sets of three or more events.

1.1.3 Conditional probability

The concepts of independence and dependence are closely related to that of 'conditional probability'. If A and B are two events with positve probabilities, then the conditional probability of A, given B, is denoted by $P(A|B)$ and is defined as

$$P(A|B) = \frac{P(AB)}{P(B)} \, . \tag{1.8}$$

If A and B are independent, then $P(A|B) = P(A)$, which is consistent with the idea that the occurrence of B does not influence the probability of A.

An intuitive justification of the definition (1.8) can be given by interpreting the probability of an event as the frequency with which that event occurs when the experiment is performed a large number of times. The ratio $P(AB)/P(B)$ can then be interpreted as the frequency of occurrence of AB among those experiments in which B occurs. Hence, that ratio is the probability that A occurs, given that B has occurred.

In terms of conditional probabilities, the joint probability that A and B occur can be expressed, according to (1.8), as

$$P(AB) = P(A|B)P(B) = P(B|A)P(A) \, . \tag{1.9}$$

This formula generalizes easily to more than two events:

$$P(A_1A_2\dots A_n) = \left[\prod_{i=1}^{n-1} P(A_i|A_{i+1}\dots A_n)\right] P(A_n)\,. \tag{1.10}$$

The probability of a given event, A, can often be determined by 'conditioning' it upon the occurrence of one of several other events. Let $B_1, B_2, \dots$ be a complete set of events, i.e. a partition (finite or infinite) of Ω. Any event, A, can be represented as

$$A = A\Omega = A\bigcup_{i=1}^{\infty} B_i = \bigcup_{i=1}^{\infty} AB_i\,, \tag{1.11}$$

where the events AB_i $(i = 1, 2, \dots)$ are disjoint. Hence,

$$P(A) = \sum_{i=1}^{\infty} P(AB_i) = \sum_{i=1}^{\infty} P(A|B_i)P(B_i)\,. \tag{1.12}$$

This expression is known as the 'complete probability formula'. It yields the probability of an arbitrary event, A, in terms of the probabilities $P(B_i)$ and the conditional probabilities $P(A|B_i)$. We shall see numerous applications of this approach.

Alternatively, having observed that the event A has occurred, one may ask what is the probability of occurrence of some B_i. This is given by what is known as the 'Bayes formula':

$$P(B_i|A) = \frac{P(B_iA)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{\infty} P(A|B_j)P(B_j)}\,. \tag{1.13}$$

Examples

5. In a group of young people consisting of 60 men and 40 women, the men divide into 20 smokers and 40 non-smokers, while the women are all non-smokers. If we know that men smokers, men non-smokers and women non-smokers survive beyond the age of 70 with probabilities 0.8, 0.85 and 0.9, respectively, what is the probability that a person chosen at random from the group will survive beyond the age of 70?

The desired quantity can be determined by applying the complete probability formula. Let A be the event 'the person chosen at random will survive beyond the age of 70'; $B1$, $B2$ and $B3$ are the events 'the person is a man smoker', 'the person is a man non-smoker' and 'the person is a woman', respectively. The three events B_1, B_2 and B_3 form a partition of Ω because one, and only one, of them occurs. Their probabilities are

$P(B_1) = 20/100 = 0.2$; $P(B_2) = 40/100 = 0.4$; $P(B_3) = 40/100 = 0.4$. Hence we can write

$$\begin{aligned} P(A) &= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3) \\ &= 0.8 \times 0.2 + 0.85 \times 0.4 + 0.9 \times 0.4 = 0.86\ . \end{aligned}$$

6. In the same group of people, suppose that event A is observed, i.e. the person chosen at random survives beyond the age of 70. What is the probability that that person is a man smoker?

Now we apply Bayes' formula:

$$\begin{aligned} P(B_1|A) &= \frac{P(A|B_1)P(B_1)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3)} \\ &= \frac{0.16}{0.86} \approx 0.186\ . \end{aligned}$$

Exercises

1. Imagine an experiment consisting of tossing a coin infinitely many times. The possible outcomes are infinite sequences of 'heads' or 'tails'. Show that, with a suitable representation of outcomes, the sample space Ω is equivalent to the closed interval $[0, 1]$.

2. For the same experiment, the event 'a head appears for the first time on the ith toss of the coin' is represented by a sub-interval of $[0, 1]$. Which sub-interval?

3. An experiment consists of attempting to compile three student programs. Each program is either accepted by the compiler as valid, or is rejected. Describe the sample space Ω. Assuming that each outcome is equally likely, find the probabilities of the following events:

A: programs 1 and 2 are accepted;
B: at least one of the programs 2 and 3 is accepted;
C: at least one of the three programs is rejected;
D: program 3 is rejected.

4. For the same experiment, show that the events A and B are dependent, as are also B and C, and C and D. However, events A and D are independent. Find the conditional probabilities $P(A|B)$, $P(B|A)$, $P(C|B)$ and $P(D|C)$.

1.2 Random variables

It is often desirable to associate various numerical values with the outcomes of an experiment, whether those outcomes are themselves numeric or not. In other words it is of interest to consider functions which are defined on a sample space Ω and whose values are real numbers. Such functions are called 'random variables'. The term 'random' refers, of course, to the fact that the value of the function is not known before the experiment is performed. After that, there is a single outcome and hence a known value. The latter is called a 'realization', or an 'instance' of the random variable.

Examples

1. A life insurance company keeps the information that it has on its customers in a large database. Suppose that a customer is selected at random. An outcome of this experiment is a collection, c, of data items describing the particular customer. The following functions of c are random variables:

$X(c) =$ 'year of birth';
$Y(c) =$ '0 if single, 1 if married';
$Z(c) =$ 'sum insured';
$V(c) =$ 'yearly premium'.

2. The lifetime of a battery powering a child's toy is measured. The sample points are now positive real numbers: $\Omega = \{x : x \in \mathcal{R}^+\}$. Those points themselves can be the values of a random variable: $Y(x) = x$.

3. The execution times, x_i, of n consecutive jobs submitted to a computer are measured. This is an experiment whose outcomes are vectors, v, with n non-negative elements: $v = (x_1, x_2, \ldots, x_n)$; $x_i \geq 0$, $i = 1, 2, \ldots, n$. Among the random variables which may be of interest in this connection are:

$X(v) = \max(x_1, x_2, \ldots, x_n)$ (longest execution time);
$Y(v) = \min(x_1, x_2, \ldots, x_n)$ (shortest execution time);
$Z(v) = (x_1 + x_2 + \ldots + x_n)/n$ (sample average execution time).

4. A function which takes a fixed value, no matter what the outcome of the experiment, e.g. $X(\omega) = 5$ for all $\omega \in \Omega$, is also a random variable, despite the fact that it is not really 'random'.

* * *