1

Elements of probability theory

1.1 Definitions

The concept of probability is of considerable importance in optics, as in any situation in which the outcome of a given trial or measurement is uncertain. Under these conditions it is desirable to be able to associate a measure with the likelihood of the outcome or the event in question; such a measure is called the *probability* of the event.

Several different definitions of probability have been adopted at various times in the past. The classical definition is based on an exhaustive enumeration of the possible outcomes of an experiment or trial. If the trial has N distinguishable, mutually exclusive outcomes, which are equally likely to occur, and if n out of these N possible outcomes have an attribute or characteristic that we call 'success', then the probability of success in any one trial is given by the ratio n/N. For example, if we roll a die, and if each of the six digits is equally likely to be on top when the die comes to rest, there are $N = 6$ distinguishable outcomes. If we identify success with an even number, for example, then since there are three different ways in which success can be achieved, it follows that the probability of success when the die is rolled is given by $3/6 = 1/2$. Unfortunately, an exhaustive enumeration of all possibilities is not always feasible.

Another common definition of probability is based on the notion of relative frequency of success. If in a large number of N independent trials the successful attribute appears n times, then the relative frequency of success is n/N. When N becomes very great, we identify this ratio with the *probability of success* in any one trial. However, n/N does not have a limit as $N \to \infty$ in the mathematical sense.

Alternatively, the concept of probability can be introduced in an axiomatic way, in which we simply associate measures $p(A)$, $p(B)$, $p(C)$, ... that we call probabilities with all possible outcomes or events A, B, C, ... of a trial. If the total event space is denoted by Ω, then $A \in \Omega$, $B \in \Omega$, etc. It is convenient to introduce the following notation which is illustrated geometrically by the Venn diagrams in Fig. 1.1:

$$A \cup B \Rightarrow A + B$$

denotes the *combination* or *union* of the two events A and B, which implies

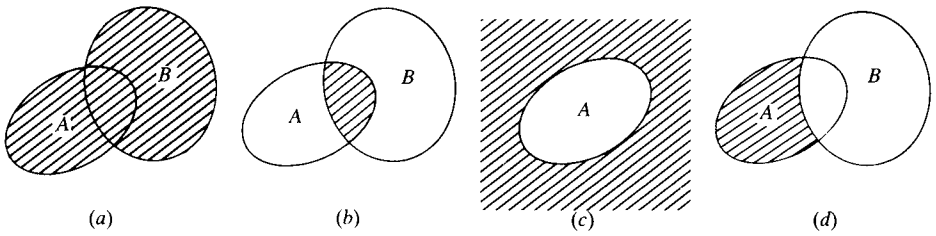

Fig. 1.1 Venn diagrams for certain combinations of events A and B. The shaded area illustrates the notion of (a) A or B or both; (b) both A and B; (c) not A; (d) A but not B.

either A, or B, or both (see Fig. 1.1(a));

$$A \cap B \Rightarrow A, B$$

denotes the *intersection* of the two events A and B, which implies both A and B (see Fig. 1.1(b));

$$\tilde{A} \Rightarrow -A$$

denotes the *complement* of the event A, which implies not A (see Fig. 1.1(c));

$$A \cap \tilde{B} \Rightarrow A - B$$

denotes the intersection of event A with the complement of B, which implies A but not B (see Fig. 1.1(d)). The null event $\varnothing$ is the complement of Ω. In all cases the notation on the right is the one customary in probability theory, and that on the left is the usual set theoretic notation. Figures 1.1(a) to 1.1(d) illustrate the notions of union of two events, intersection of two events, etc., etc., geometrically. Two events A and B are said to be *disjoint* or *mutually exclusive* if they do not overlap at all, or the intersection A, B is the null event $\varnothing$.

The following three axioms suffice to determine the properties of the probability $p(A)$ of a given event:

(a) $p(A) \geqslant 0$, (1.1–1)

(b) $p(\Omega) = 1$, (1.1–2)

(c) if $A_1, A_2, A_3, \ldots$ are mutually exclusive events, then

$$p(A_1 + A_2 + A_3 + \ldots) = p(A_1) + p(A_2) + p(A_3) + \ldots . \quad (1.1–3)$$

Equation (1.1–2) may be interpreted to mean that the probability of an outcome that is certain is unity. As $A + \tilde{A} = \Omega$, and A and $\tilde{A}$ are mutually exclusive, it follows from Eq. (1.1–3) that

$$p(A) + p(\tilde{A}) = p(A + \tilde{A}) = p(\Omega),$$

and from Eqs. (1.1–1) and (1.1–2)

$$0 \leqslant p(A) \leqslant 1. \quad (1.1–4)$$

The bounds on any probability are therefore zero from below and unity from above.

1.2 Properties of probabilities

Several important corollaries follow immediately from these relations. If the events $A_1, A_2, A_3, \ldots, A_N$ are mutually exclusive and represent the set of all possible outcomes, so that $A_1 + A_2 + \ldots + A_N = \Omega$, then from Eqs. (1.1–2) and (1.1–3) it follows that

$$\sum_{i=1}^{N} p(A_i) = p\left(\sum_{i=1}^{N} A_i\right) = p(\Omega) = 1. \tag{1.2–1}$$

Also, if A is a subset of B, or $A \subset B$, as illustrated by the Venn diagram in Fig. 1.2, then A and $B - A$ are mutually exclusive, and their union is B, so that by Eq. (1.1–3)

$$p(A) + p(B - A) = p(B),$$

or

$$p(A) \leqslant p(B) \quad \text{when } A \subset B. \tag{1.2–2}$$

Finally, we note that an event that cannot occur has probability zero, because $\varnothing + \Omega = \Omega$, so that

$$p(\varnothing) + p(\Omega) = p(\Omega),$$

and

$$p(\varnothing) = 0. \tag{1.2–3}$$

Thus if A and B are mutually exclusive, then the probability of both A and B $p(A, B) = 0$.

1.2.1 Joint probabilities

Events that are obtained by compounding other events are known as joint events, and the corresponding probabilities are joint probabilities. Thus $p(A, B)$ is the *joint probability* of both events A and B, or the probability of the intersection of A with B. The order in which the events A and B are listed is immaterial. As the compound event A, B is a subset of the event A (see Fig. 1.1(b)), it follows from Eq. (1.2–2) that

$$\left.\begin{aligned} &p(A, B) \leqslant p(A), \\ \text{and similarly} \quad &p(A, B) \leqslant p(B). \end{aligned}\right\} \tag{1.2–4}$$

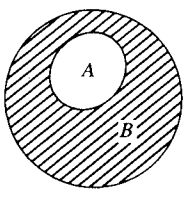

Fig. 1.2 Illustrating A as a subset of B.

The joint probability for two events is therefore always less than or equal to the probability for one of the events alone.

If $B_1, B_2, \ldots, B_M$ is a set of all possible mutually exclusive events, then

$$\sum_{i=1}^{M} p(A, B_i) = p(A, \Omega) = p(A). \tag{1.2–5}$$

This result follows immediately from the fact that $A, B_1\ A, B_2 \ldots$ is also a set of mutually exclusive events spanning the whole space (see Eq. (1.1–3)). More generally, joint probabilities may involve more than two events. If $C_1, C_2, \ldots, C_N$ is a complete set of mutually exclusive events, then

$$\sum_{k=1}^{N} p(A, B, C_k, D, \ldots) = p(A, B, D, \ldots). \tag{1.2–6}$$

Let us now consider the situation in which two events A and B are not necessarily mutually exclusive (see Fig. 1.1(a)), and let us calculate the probability $p(A + B)$ of the union $A + B$. We cannot apply the summation law (1.2–6) to A and B directly. However, we note that the two events A and $B - A$ (cf. Figs. 1.1) are mutually exclusive, and their union is $A + B$. Then according to Eq. (1.1–3),

$$p(A) + p(B - A) = p(A + B). \tag{1.2–7}$$

Also, $B - A$ and A, B are mutually exclusive with union B, so that

$$p(B) = p(B - A) + p(A, B).$$

If we substitute for $p(B - A)$ in Eq. (1.2–7), we obtain immediately

$$p(A + B) = p(A) + p(B) - p(A, B), \tag{1.2–8}$$

so that

$$p(A + B) \leqslant p(A) + p(B). \tag{1.2–9}$$

Equation (1.2–8) is known as the *composition law* for two events that are not necessarily mutually exclusive. The relation is readily generalized to N events $A_1, A_2, \ldots, A_N$, for which it may be proved by induction that

$$p(A_1 + A_2 + \ldots + A_N) = \sum_{i=1}^{N} p(A_i) - \underbrace{\sum_{i \neq j}^{N}\sum^{N} p(A_i, A_j)}_{\binom{N}{2}\text{ pairs}} + \underbrace{\sum_{\substack{i \neq j \neq k \\ i \neq k}}^{N}\sum^{N}\sum^{N} p(A_i, A_j, A_k)}_{\binom{N}{3}\text{ triplets}}$$
$$- \ldots + (-1)^{N-1} p(A_1, A_2, \ldots, A_N). \tag{1.2–10}$$

Also, by repeated application of the inequality (1.2–9), one readily finds that

$$\begin{aligned} p(A_1 + A_2 + \ldots + A_N) &\leqslant p(A_1 + A_2 + \ldots + A_{N-1}) + p(A_N) \\ &\leqslant p(A_1 + A_2 + \ldots + A_{N-2}) + p(A_{N-1}) + p(A_N) \\ &\leqslant p(A_1) + p(A_2) + \ldots + p(A_N), \end{aligned} \tag{1.2–11}$$

with the equality sign holding for the special case of mutually exclusive events.

1.2.2 *Conditional probabilities*

The probability of some event A conditioned on some other event B is known as the *conditional probability of A given B*, and it is frequently denoted by $\mathcal{P}(A|B)$. It is given by the ratio

$$\mathcal{P}(A|B) = p(A, B)/p(B), \tag{1.2–12}$$

and it is, of course, defined only when B is not a null event. From Eq. (1.2–4) it follows immediately that

$$0 \leqslant \mathcal{P}(A|B) \leqslant 1, \tag{1.2–13}$$

so that a conditional probability is a true probability, with all the properties given earlier. If $A_1, A_2, \ldots, A_N$ is a complete set of mutually exclusive possible outcomes, then by virtue of the property (1.2–5) we have

$$\sum_{i=1}^{N} \mathcal{P}(A_i|B) = 1. \tag{1.2–14}$$

If the conditional probability of A given B is equal to the unconditional probability of A, or

$$\mathcal{P}(A|B) = p(A), \tag{1.2–15}$$

then it evidently does not matter whether event B occurs, or not, so far as event A is concerned. Events A and B are then described as being *statistically independent*. From Eqs. (1.2–15) and (1.2–12) we see that

$$p(A, B) = p(A)p(B) \tag{1.2–16}$$

whenever A and B are statistically independent, and this is sometimes taken to be the defining relation for statistical independence. More generally, the necessary and sufficient condition for N events $A_1, A_2, \ldots, A_N$ to be statistically independent is that the joint probability factorizes in the form

$$p(A_1, A_2, \ldots, A_N) = p(A_1)p(A_2) \ldots p(A_N). \tag{1.2–17}$$

A similar relation then holds for any subset of the N events. Needless to say, events that are mutually exclusive cannot be statistically independent, because the joint probability for mutually exclusive events is zero (except for the trivial case in which one or more of the events cannot happen at all).

A simple example may be helpful. Suppose that a die is rolled, and the number ending up on top is registered. We are interested in events of type A in which the number is divisible by 2, events B in which the number is divisible by 3, and events C in which the number is prime. These events are described by the following sets, with the indicated probabilities:

$$\left.\begin{aligned} A &= (2, 4, 6), \quad p(A) = \tfrac{1}{2} \\ B &= (3, 6), \quad\;\; p(B) = \tfrac{1}{3} \\ C &= (2, 3, 5), \quad p(C) = \tfrac{1}{2}. \end{aligned}\right\} \tag{1.2–18}$$

The intersections among these sets are given by

$$\left.\begin{aligned}(A, B) &= (6), \quad p(A, B) = \tfrac{1}{6}\\ (A, C) &= (2), \quad p(A, C) = \tfrac{1}{6}\\ (B, C) &= (3), \quad p(B, C) = \tfrac{1}{6}.\end{aligned}\right\} \tag{1.2–19}$$

It follows that

$$\left.\begin{aligned}p(A, B) &= p(A)p(B)\\ p(A, C) &\neq p(A)p(C)\\ p(B, C) &= p(B)p(C),\end{aligned}\right\} \tag{1.2–20}$$

so that A and B are statistically independent, as are B and C, but A and C are not statistically independent.

1.2.3 Bayes' theorem on inverse probabilities

From the definition (1.2–12) of conditional probability the following two relations follow:

$$p(A, B) = \mathcal{P}'(A|B)p(B)$$
$$p(A, B) = \mathcal{P}(B|A)p(A).$$

On equating both expressions for $p(A, B)$ we obtain for mutually exclusive events A and B

$$\mathcal{P}'(A|B) = \frac{\mathcal{P}(B|A)p(A)}{p(B)} = \frac{\mathcal{P}(B|A)p(A)}{\sum\limits_{\text{all}\,A} \mathcal{P}(B|A)p(A)}, \tag{1.2–21}$$

where we have made use of Eq. (1.2–5) in the last expression. This relation is known as *Bayes' theorem*. If we call $\mathcal{P}(B|A)$ the conditional probability of B given A, we may think of $\mathcal{P}'(A|B)$ as the *inverse probability* of A given B. Bayes' theorem then allows the inverse probability to be determined from the forward conditional probability together with $p(A)$. In practice the theorem is often applied to experimental situations in which A is to be determined from measurements of B, but little or nothing is known about the a priori probability $p(A)$. Some assumption about the a priori characteristics of $p(A)$ then has to be made before Eq. (1.2–21) can be used, and this introduces a certain arbitrariness into the procedure, which has been criticized.

Let us illustrate the problem by a simple example. A vessel contains N balls, which are either black or white, in unknown proportion. A ball is picked at random and is found to be white. We wish to determine the inverse probability that the vessel contained n ($0 \leqslant n \leqslant N$) white balls originally in the light of the experiment. Let $\mathcal{P}(1|n)$ be the conditional probability that a white ball is picked when the vessel actually contains n white balls ($n = 0, 1, \ldots, N$). From the nature of the problem it is evident that $\mathcal{P}(1|n) = n/N$. Then from Eq. (1.2–21) the inverse probability $\mathcal{P}'(n|1)$ that the vessel originally contained n white balls,

given that a white ball is picked, is given by

$$\mathscr{P}'(n|1) = \frac{\mathscr{P}(1|n)p(n)}{\sum_{n=0}^{N} \mathscr{P}(1|n)p(n)} = \frac{(n/N)p(n)}{\sum_{n=0}^{N} (n/N)p(n)}, \tag{1.2–22}$$

where $p(n)$ is the a priori probability that the vessel contains n white balls. Unfortunately, nothing is known about $p(n)$, so that, strictly speaking, Eq. (1.2–22) cannot be applied. However, in the absence of further information, if we arbitrarily assign equal weights to all values of n from 0 to N a priori, then $p(n) = 1/(N+1)$ and Eq. (1.2–22) leads to the solution

$$\mathscr{P}'(n|1) = \frac{n}{\sum_{n=0}^{N} n} = \frac{n}{\frac{1}{2}N(N+1)}. \tag{1.2–23}$$

By making some assignment to the a priori probabilities $p(n)$, we have been able to calculate the inverse probabilities $\mathscr{P}'(n|1)$. Although it may not be possible to give a formal justification for this procedure, it nevertheless leads to quantitative estimates that are often valuable.

1.3 Random variables and probability distributions

When the possible outcomes A of a trial or experiment are numbers, then the outcomes are automatically mutually exclusive. It is convenient to regard these numbers as the values of some variable x, which is known as a *random variable* or *variate*. If the possible values of x consist of the countable set of numbers x_1, x_2, x_3, ..., then x is known as a *discrete random variable*, whereas if the possible values are any numbers in some interval (a, b) (which may be infinite), x is known as a *continuous random variable*. The set of all possible outcomes is known as the *ensemble* of x. Usually the random variable is taken to be real, but complex random variables $z = x + iy$, whose real and imaginary parts x, y are both random variables, will also be encountered.

With each of the possible outcomes or values x_1, x_2, ... of the discrete variate we may associate a probability p_i ($i = 1, 2, \ldots$), and as the different values are mutually exclusive, the corresponding probabilities must sum to unity, by virtue of Eq. (1.2–1),

$$\sum_{\text{all } i} p_i = 1. \tag{1.3–1}$$

A graph of the probability p_i versus x_i as in Fig. 1.3(a) consists of a series of points or lines, and illustrates the distribution of probability over the interval. In the special case in which one value x_0 is certain, and none of the other values x_1, x_2, ... occurs, the form of p_i is

$$p_i = \delta_{i0}, \tag{1.3–2}$$

where δ_{ij} is the Kronecker delta symbol, i.e. $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise.

If x is a continuous variate in the interval (a, b), it is convenient to associate a

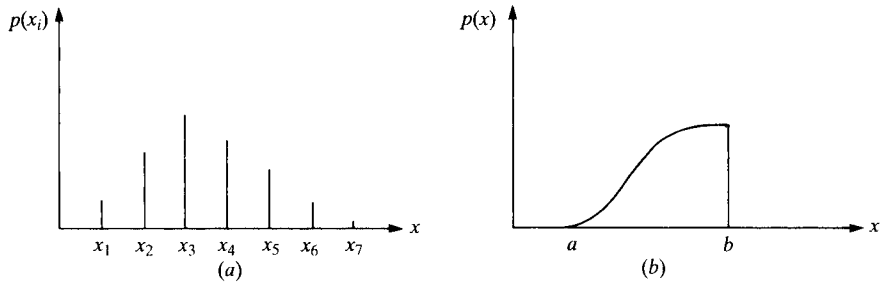

Fig. 1.3 Illustrating (*a*) discrete, (*b*) continuous probability distributions.

probability density $p(x)$ with the ensemble of x, such that $p(x)\,\mathrm{d}x$ gives the probability for x to be found in the infinitesimal interval from x to $x + \mathrm{d}x$. Then, corresponding to Eq. (1.3–1), we have the normalization condition

$$\int_a^b p(x)\,\mathrm{d}x = 1. \tag{1.3–3}$$

The form of $p(x)$ gives the probability distribution of the variate x (see Fig. 1.3(*b*)). The probability $P(x \leqslant X)$ for x to be equal to or below X ($a \leqslant X \leqslant b$) is given by the integral

$$P(x \leqslant X) = \int_a^X p(x)\,\mathrm{d}x. \tag{1.3–4}$$

Corresponding to Eq. (1.2–2) we have the relation

$$P(x \leqslant X_1) \leqslant P(x \leqslant X_2) \quad \text{when } X_1 \leqslant X_2. \tag{1.3–5}$$

$P(x \leqslant X)$ is therefore an increasing function of X that is bounded by unity, and its derivative is the probability density

$$\frac{\mathrm{d}P(x \leqslant X)}{\mathrm{d}X} = p(X). \tag{1.3–6}$$

The probability density $p(X)$ may not exist as an ordinary function when $P(x \leqslant X)$ is discontinuous, but it cannot be more singular than a Dirac delta function. If the continuous random variable x takes on the value x_0 with certainty, then $p(x)$ has the form

$$p(x) = \delta(x - x_0), \tag{1.3–7}$$

which can be compared with Eq. (1.3–2) for a discrete random variable. The need for delta functions to describe probability densities can be avoided by the use of Stieltjes integrals (Yaglom, 1962, Chap. 2, Sec. 9), but we shall not hesitate to use delta functions here. Indeed, by the introduction of delta functions we can incorporate the treatment of discrete variates in the treatment of continuous variates. If a discrete variate takes on the values $x_1, x_2, \ldots$ with probabilities p_1, $p_2, \ldots$, then we can formally describe this situation by a continuous variate x having the following probability density $p(x)$:

$$p(x) = \sum_i p_i \delta(x - x_i). \tag{1.3–8}$$

Because of this, and in order to avoid repetition, we shall henceforth formally regard x as a continuous variate.

A cautionary note regarding notation may be in order here. If x and y are two different random variables, their probability densities are sometimes denoted by $p(x)$ and $p(y)$, respectively, without any implication that the functional forms of the two probability distributions are equal. However, it is generally safer to use different symbols, e.g. $p(x)$ and $P(y)$, for two probability densities that are not necessarily equal.

1.3.1 Transformations of variates

Let x be a random variable defined on the interval (a, b) with probability density $p(x)$. It is sometimes necessary to make a transformation from x to a new variable y, where

$$y = f(x), \quad A \leqslant y \leqslant B, \tag{1.3–9}$$

and we wish to determine the probability density $P(y)$ of y. Let us first suppose that the transformation (1.3–9) has a single-valued inverse,

$$x = g(y). \tag{1.3–10}$$

Then if x and y correspond to each other, and the interval $\mathrm{d}x$ corresponds to the interval $\mathrm{d}y$, then evidently

$$P(y)|\mathrm{d}y| = p(x)|\mathrm{d}x|,$$

so that

$$\begin{aligned} P(y) &= p(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| \\ &= p[g(y)]|g'(y)| \\ &= \frac{p[g(y)]}{|f'[g(y)]|}. \end{aligned} \tag{1.3–11}$$

More generally, if the inverse is multivalued, and to a given y there correspond several x's

$$\left.\begin{aligned} x_1 &= g_1(y) \\ x_2 &= g_2(y) \\ &\cdots\cdots, \end{aligned}\right\} \tag{1.3–12}$$

then we need to add the probabilities associated with these different, mutually exclusive x's, and we have in place of Eq. (1.3–11)

$$\begin{aligned} P(y) &= \sum_i p(x_i)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right|_{x=x_i} \\ &= \sum_i p[g_i(y)]|g_i'(y)| \\ &= \sum_i \frac{p[g_i(y)]}{|f'[g_i(y)]|}. \end{aligned} \tag{1.3–13}$$

The same result can also be formally expressed in the more compact form

$$P(y) = \int p(x)\delta[y - f(x)]\,\mathrm{d}x, \tag{1.3–14}$$

if we expand the delta function in the usual way in terms of its zeros,

$$\delta[y - f(x)] = \sum_i \frac{\delta(x - x_i)}{|f'(x_i)|}. \tag{1.3–15}$$

Equation (1.3–14) can be interpreted to mean that the probability density of y is obtained by integrating the probability density $p(x)$ of x over all those values of x that correspond to y, i.e. those which are subject to the constraint $y = f(x)$.

As an example we consider the change of probability under the transformation $y = x^2$, which has the double-valued inverse

$$x = \pm\sqrt{y}$$

$$\frac{\mathrm{d}x}{\mathrm{d}y} = \pm\frac{1}{2\sqrt{y}}.$$

In this case Eq. (1.3–14) gives

$$P(y) = \frac{p(\sqrt{y})}{2\sqrt{y}} + \frac{p(-\sqrt{y})}{2\sqrt{y}}, \quad 0 \leqslant y. \tag{1.3–16}$$

Next we consider the more general situation in which we have a set of N variates $x_1, x_2, \ldots, x_N$, with joint probability density $p(x_1, x_2, \ldots, x_N)$, and we wish to transform to a new set $y_1, y_2, \ldots, y_N$, with probability density $P(y_1, y_2, \ldots, y_N)$. If the transformation

$$y_r = f_r(x_1, x_2, \ldots, x_N), \quad (r = 1, 2, \ldots, N)$$

has a single-valued inverse

$$x_r = g_r(y_1, y_2, \ldots, y_N), \quad (r = 1, 2, \ldots, N),$$

then

$$P(y_1, y_2, \ldots, y_N)|\mathrm{d}y_1\,\mathrm{d}y_2 \ldots \mathrm{d}y_N| = p(x_1, x_2, \ldots, x_N)|\,\mathrm{d}x_1\,\mathrm{d}x_2 \ldots \mathrm{d}x_N|, \tag{1.3–17a}$$

and

$$P(y_1, y_2, \ldots, y_N) = |J|p(x_1, x_2, \ldots, x_N), \tag{1.3–17b}$$

where J is the Jacobian of the transformation

$$|J| = \left|\frac{\partial(g_1, g_2, \ldots, g_N)}{\partial(y_1, y_2, \ldots, y_N)}\right|.$$

Once again we can express the transformation of the probability with the help of