2. Probability Concepts

In the preceding chapter, we introduced probability notions without any definitions. In order to formulate essential concepts more precisely, it is necessary to have some more precise expression of these concepts. The intention of this chapter is to provide some background, and to present a number of essential results. It is not a thorough outline of mathematical probability, for which the reader is referred to standard mathematical texts such as those by *Feller* [2.1] and *Papoulis* [2.2].

2.1 Events, and Sets of Events

It is convenient to use a notation which is as general as possible in order to describe those occurrences to which we might wish to assign probabilities. For example, we may wish to talk about a situation in which there are 6.4×10^{14} molecules in a certain region of space; or a situation in which a Brownian particle is at a certain point $\boldsymbol{x}$ in space; or possibly there are 10 mice and 3 owls in a certain region of a forest.

These occurrences are all examples of practical realisations of *events*. More abstractly, an event is simply a member of a certain space, which in the cases most practically occurring can be characterised by a vector of integers

$$\boldsymbol{n} = (n_1,\ n_2,\ n_3\ \dots), \tag{2.1.1}$$

or a vector of real numbers

$$\boldsymbol{x} = (x_1,\ x_2,\ x_3\ \dots). \tag{2.1.2}$$

The dimension of the vector is arbitrary.

It is convenient to use the language of set theory, introduce the concept of a *set of events*, and use the notation

$$\omega \in A, \tag{2.1.3}$$

to indicate that the event ω is one of events contained in A. For example, one may consider the set $A(25)$ of events in the ecological population in which there are no more than 25 animals present; clearly the event $\bar{\omega}$ that there are 3 mice, a tiger, and no other animals present satisfies

$$\bar{\omega} \in A(25). \tag{2.1.4}$$

More significantly, suppose we define the set of events $A(\boldsymbol{r}, \Delta V)$ that a molecule is within a volume element ΔV centred on a point $\boldsymbol{r}$. In this case, the practical significance of working in terms of sets of events becomes clear, because we should normally be able to determine whether or not a molecule is within a neighbourhood ΔV of $\boldsymbol{r}$, but to determine whether the particle is exactly at $\boldsymbol{r}$ is impossible. Thus, if we define the event $\omega(\boldsymbol{y})$ that the molecule is at point $\boldsymbol{y}$, it makes sense to ask whether

$$\omega(\boldsymbol{y}) \in A(\boldsymbol{r}, \Delta V)\,, \tag{2.1.5}$$

and to assign a certain probability to the *set* $A(\boldsymbol{r}, \Delta V)$, which is to be interpreted as the probability of the occurrence of (2.1.5).

2.2 Probabilities

Most people have an intuitive conception of a probability, based on their own experience. However, a precise formulation of intuitive concepts is fraught with difficulties, and it has been found most convenient to axiomatise probability theory as an essentially abstract science, in which a probability measure $P(A)$ is *assigned* to every set A, in the space of events, including

$$\text{The set of all events } : \varOmega\,, \tag{2.2.1}$$

$$\text{The set of no events } : \varnothing\,, \tag{2.2.2}$$

in order to define probability, we need our sets of events to form a closed system (known by mathematicians as a *σ-algebra*) under the set theoretic operations of union and intersection.

2.2.1 Probability Axioms

We introduce the probability of A, $P(A)$, as a function of A satisfying the following *probability axioms*:

i) $P(A) \geqslant 0 \quad \text{for all } A\,,$ (2.2.3)

ii) $P(\varOmega) = 1\,,$ (2.2.4)

iii) If $A_i \quad (i = 1, 2, 3, \ldots)$ is a countable (but possibly infinite) collection of nonoverlapping sets, i.e., such that

$$A_i \cap A_i = \varnothing \quad \text{for all} \quad i \neq j\,, \tag{2.2.5}$$

then

$$P\Big(\bigcup_i A\Big) = \sum_i P(A_i)\,. \tag{2.2.6}$$

These are all the axioms needed. Consequentially, however, we have:

iv) if $\bar{A}$ is the complement of A, i.e., the set of all events not contained in A, then

$$P(\bar{A}) = 1 - P(A)\,, \tag{2.2.7}$$

v) $P(\varnothing) = 0\,.$ (2.2.8)

2.2.2 The Meaning of $P(A)$

There is no way of making probability theory correspond to reality without requiring a certain degree of intuition. The probability $P(A)$, as axiomatised above, is the intuitive probability that an "*arbitrary*" event ω, i.e., an event ω "*chosen at random*", will satisfy $\omega \in A$. Or more explicitly, if we choose an event "*at random*" from Ω N times, the relative frequency that the particular event chosen will satisfy $\omega \in A$ approaches $P(A)$ as the number of times, N, we choose the event, approaches infinity. The number of choices N can be visualised as being done one after the other ("*independent*" tosses of one die) or at the same time (N dice are thrown at the same time "*independently*"). All definitions of this kind must be intuitive, as we can see by the way undefined terms ("*arbitrary*", "*at random*", "*independent*") keep turning up. By eliminating what we now think of as intuitive ideas and axiomatising probability, *Kolmogorov* [2.3] cleared the road for a rigorous development of mathematical probability. But the circular definition problems posed by wanting an intuitive understanding remain. The simplest way of looking at axiomatic probability is as a formal method of manipulating probabilities using the axioms. In order to apply the theory, the probability space must be defined *and* the probability measure P assigned. These are *a priori probabilities*, which are simply assumed. Examples of such a priori probabilities abound in applied disciplines. For example, in equilibrium statistical mechanics one assigns equal probabilities to equal volumes of phase space. Einstein's reasoning in Brownian motion assigned a probability $\phi(\Delta)$ to the probability of a "push" Δ from a position x at time t.

The task of applying probability is

i) To assume some set of *a priori* probabilities which seem reasonable and to deduce results from this and from the structure of the probability space,

ii) To measure experimental results with some apparatus which is constructed to measure quantities in accordance with these a priori probabilities.

The structure of the probability space is very important, especially when the space of events is compounded by the additional concept of time. This extension makes the effective probability space infinite-dimensional, since we can construct events such as "the particle was at points $\boldsymbol{x}_n$ at times t_n for $n = 0, 1, 2, \ldots, \infty$".

2.2.3 The Meaning of the Axioms

Any intuitive concept of probability gives rise to nonnegative probabilities, and the probability that an arbitrary event is contained in the set of all events must be 1 no matter what our definition of the word arbitrary. Hence, axioms i) and ii) are understandable. The heart of the matter lies in axiom iii). Suppose we are dealing with only 2 sets A and B, and $A \cap B = \emptyset$. This means there are *no* events contained in both A and B. Therefore, the probability that $\omega \in A \cup B$ is the probability that *either* $\omega \in A$ or $\omega \in B$. Intuitive considerations tell us this probability is the sum of the individual probabilities, i.e.,

$$P(A \cup B) \equiv P\{(\omega \in A) \text{ or } (\omega \in B)\} = P(A) + P(B). \tag{2.2.9}$$

Notice this is not a proof—merely an explanation.

The extension now to any finite number of nonoverlapping sets is obvious, but the extension only to any *countable* number of nonoverlapping sets requires some comment.

This extension must be made restrictive because of the existence of sets labelled by a continuous index, for example, $\boldsymbol{x}$, the position in space. The probability of a molecule being in the set whose only element in $\boldsymbol{x}$ is zero; but the probability of being in a region R of finite volume is nonzero. The region R is a union of sets of the form $\{\boldsymbol{x}\}$—but not a *countable* union. Thus axiom iii) is not applicable and the probability of being in R is *not* equal to the sum of the probabilities of being in $\{\boldsymbol{x}\}$.

2.2.4 Random Variables

The concept of a random variable is a notational convenience which is central to this book. Suppose we have an abstract probability space whose events can be written $\boldsymbol{x}$. Then we can introduce the random variable $F(\boldsymbol{x})$ which is a function of $\boldsymbol{x}$, which takes on certain values for each $\boldsymbol{x}$. In particular, the identity function of $\boldsymbol{x}$, written $\boldsymbol{X}(\boldsymbol{x})$ is of interest; it is given by

$$X(\boldsymbol{x}) = \boldsymbol{x}. \tag{2.2.10}$$

We shall normally use capitals in this book to denote random variables and small letters $\boldsymbol{x}$ to denote their values whenever it is necessary to make a distinction.

Very often, we have some quite different underlying probability space Ω with values ω, and talk about $\boldsymbol{X}(\omega)$ which is some function of ω, and then omit explicit mention of ω. This can be for either of two reasons:

i) we specify the events by the values of $\boldsymbol{x}$ anyway, i.e., we identify $\boldsymbol{x}$ and ω;
ii) the underlying events ω are too complicated to describe, or sometimes, even to know.

For example, in the case of the position of a molecule in a liquid, we really should interpret each ω as being capable of specifying all the positions, momenta, and orientations of each molecule in that volume of liquid; but this is simply too difficult to write down, and often unnecessary.

One great advantage of introducing the concept of a random variable is the simplicity with which one may handle functions of random variables, e.g., $\boldsymbol{X}^2$, $\sin(\boldsymbol{a}\cdot\boldsymbol{X})$, etc., and compute means and distributions of these. Further, by defining stochastic differential equations, one can also quite simply talk about time development of random variables in a way which is quite analogous to the classical description by means of differential equations of non-probabilistic systems.

2.3 Joint and Conditional Probabilities: Independence

2.3.1 Joint Probabilities

We explained in Sect. 2.2.3 how the occurrence of mutually exclusive events is related to the concept of nonintersecting sets. We now consider the concept $P(A \cap B)$, where $A \cap B$ is nonempty. An event ω which satisfies $\omega \in A$ will only satisfy $\omega \in A \cap B$ if $\omega \in B$ as well.

$$\text{Thus, } P(A \cap B) = P\{(\omega \in A) \text{ and } (\omega \in B)\}\,, \tag{2.3.1}$$

and $P(A \cap B)$ is called the *joint probability* that the event ω is contained in both classes, or, alternatively, that both the events $\omega \in A$ and $\omega \in B$ occur. Joint probabilities occur naturally in the context of this book in two ways:

i) *When the event is specified by a vector*, e.g., m mice and n tigers. The probability of this event is the joint probability of [m mice (and any number of tigers)] and [n tigers (and any number of mice)]. All vector specifications are implicitly joint probabilities in this sense.
ii) *When more than one time is considered*: what is the probability that (at time t_1 there are m_1 tigers and n_1 mice) and (at time t_2 there are m_2 tigers and n_2 mice). To consider such a probability, we have effectively created out of the events at time t_1 and events at time t_2, *joint events* involving one event at each time. In essence, there is no difference between these two cases except for the fundamental dynamical role of time.

2.3.2 Conditional Probabilities

We may specify conditions on the events we are interested in and consider only these, e.g., the probability of 21 buffaloes given that we know there are 100 lions. What does this mean? Clearly, we will be interested only in those events contained in the set $B =$ {all events where exactly 100 lions occur}. This means that we to define conditional probabilities, which are defined only on the collection of all sets contained in B. we define the conditional probability as

$$P(A\,|\,B) = P(A \cap B)/P(B)\,, \tag{2.3.2}$$

and this satisfies our intuitive conception that the conditional probability that $\omega \in A$ (given that we know $\omega \in B$), is given by dividing the probability of joint occurrence by the probability ($\omega \in B$).

We can define in both directions, i.e., we have

$$P(A \cap B) = P(A\,|\,B)P(B) = P(B\,|\,A)P(A)\,. \tag{2.3.3}$$

There is no particular conceptual difference between, say, the probability of {(21 buffaloes) given (100 lions)} and the reversed concept. However, when two times are involved, we do see a difference. For example, the probability that a particle is at position $\boldsymbol{x}_1$ at time t_1, given that it was at $\boldsymbol{x}_2$ at the *previous* time t_2, is a very natural thing to consider; indeed, it will turn out to be a central concept in this book.

The converse looks to the past rather than the future; given that a particle is at x_1 at time t_1, what is the probability that that at the previous time t_2 it was at position x_2. The first concept—the *forward* probability—looks at where the particle will go, the second—the *backward* probability—at where it came from.

The forward probability has already occurred in this book, for example, the $\phi(\Delta)d\Delta$ of Einstein (Sect. 1.2.1) is the probability that a particle at x at time t will be in the range $[x+\Delta, x+\Delta+d\Delta]$ at time $t+\tau$, and similarly in the other examples. Our intuition tells us as it told Einstein (as can be seen by reading the extract from his paper) that this kind of conditional probability is directly related to the time development of a probabilistic system.

2.3.3 Relationship Between Joint Probabilities of Different Orders

Suppose we have a collection of sets B_i such that

$$B_i \cap B_j = \varnothing\,, \tag{2.3.4}$$

$$\bigcup_i B_i = \Omega\,, \tag{2.3.5}$$

so that the sets divide up the space Ω into nonoverlapping subsets.
Then

$$\bigcup_i (A \cap B_i) = A \cap \Big(\bigcup_i B_i\Big) = A \cap \Omega = A\,. \tag{2.3.6}$$

Using now the probability axiom iii), we see that $A \cap B_i$ satisfy the conditions on the A_i used there, so that

$$\sum_i P(A \cap B_i) = P(\bigcup_i (A \cap B_i))\,, \tag{2.3.7}$$

$$= P(A)\,, \tag{2.3.8}$$

and thus

$$\sum_i P(A\,|\,B_i)P(B_i) = P(A)\,. \tag{2.3.9}$$

Thus, summing over all mutually exclusive possibilities of B in the joint probability eliminates that variable.

Hence, in general,

$$\sum_i P(A_i \cap B_j \cap C_k\ \ldots) = P(B_j \cap C_k \cap\ \ldots)\,. \tag{2.3.10}$$

The result (2.3.9) has very significant consequences in the development of the theory of stochastic processes, which depends heavily on joint probabilities.

2.3.4 Independence

We need a probabilistic way of specifying what we mean by independent events. Two sets of events A and B should represent independent sets of events if the specification that a particular event is contained in B has no influence on the probability of that event belonging to A. Thus, the conditional probability $P(A\,|\,B)$ should be independent of B, and hence

$$P(A \cap B) = P(A)P(B)\,. \tag{2.3.11}$$

In the case of several events, we need a somewhat stronger specification. The events $(\omega \in A_i)(i = 1, 2, \ldots, n)$ will be considered to be independent if for any subset $(i_1, i_2, \ldots, i_k)$ of the set $(1, 2, \ldots, n)$,

$$P(A_{i_1} \cap A_{i_2} \ldots A_{i_k}) = P(A_{i_1})P(A_{i_2}) \ldots P(A_{i_k})\,. \tag{2.3.12}$$

It is important to require factorisation for all possible combinations, as in (2.3.12). For example, for three sets A_i, it is quite conceivable that

$$P(A_i \cap A_j) = P(A_i)P(A_j)\,, \tag{2.3.13}$$

for all different i and j, but also that

$$A_1 \cap A_2 = A_2 \cap A_3 = A_3 \cap A_1\,. \qquad \text{(see Fig. 2.1)} \tag{2.3.14}$$

This requires

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3) &= P(A_2 \cap A_3 \cap A_3) = P(A_2 \cap A_3) \\ &= P(A_2)P(A_3) \neq P(A_1)P(A_2)P(A_3)\,. \end{aligned} \tag{2.3.15}$$

We can see that the occurrence of $\omega \in A_2$ and $\omega \in A_3$ necessarily implies the occurrence of $\omega \in A_1$. In this sense the events are obviously not independent.

Random variables $X_1, X_2, X_3, \ldots$, will be said to be independent random variables, if for all sets of the form $A_i = x$ such that $a_i \leqslant x \leqslant b_i$) the events $X_1 \in A_1, X_2 \in A_2, X_3 \in A_3, \ldots$ are independent events. This will mean that all values of the X_i are assumed independently of those of the remaining X_i.

2.4 Mean Values and Probability Density

The mean value (or *expectation*) of a random variable $R(\omega)$ in which the basic events ω are countably specifiable is given by

$$\langle R \rangle = \sum_{\omega} P(\omega)R(\omega)\,, \tag{2.4.1}$$

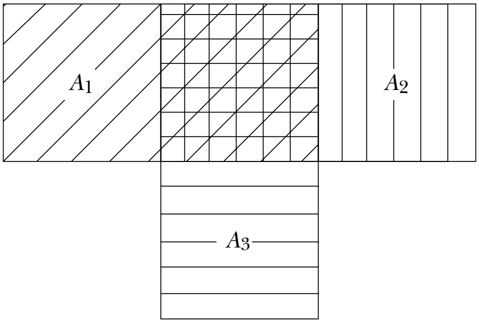

Fig. 2.1. Illustration of statistical independence in pairs, but not in threes. In the three sets $A_j \cap A_i$ is, in all cases, the central region. By appropriate choice of probabilities, it can be arranged that $P(A_i \cap A_j) = P(A_i)P(A_j)$.

where $P(\omega)$ means the probability of the set containing only the single event ω. In the case of a continuous variable, the probability axioms above enable us to define a probability density function $p(\omega)$ such that if $A(\omega_0, d\omega_0)$ is the set

$$(\omega_0 \leqslant \omega < \omega_0 + d\omega_0)\,, \tag{2.4.2}$$

then

$$p(\omega_0)d\omega_0 = P[A(\omega_0, d\omega_0)] \equiv p(\omega_0, d\omega_0)\,. \tag{2.4.3}$$

The last is a notation often used by mathematicians. Details of how this is done have been nicely explained by *Feller* [2.1]. In this case,

$$\langle R \rangle = \int_{\omega\in\Omega} d\omega\, R(\omega)p(\omega)\,. \tag{2.4.4}$$

One can often (as mentioned in Sect. 2.2.4) use R itself to specify the event, so we will often write

$$\langle R \rangle = \int dR\, R\, p(R)\,. \tag{2.4.5}$$

Obviously, $p(R)$ is not the same function of R as $p(\omega)$ is of ω—more precisely

$$p(R_0)\, dR_0 = P\,(R_0 < R < R_0 + dR_0)\ . \tag{2.4.6}$$

2.4.1 Determination of Probability Density by Means of Arbitrary Functions

Suppose for every function $f(R)$ we know

$$\langle f(R) \rangle = \int dR\, f(R)p(R)\,, \tag{2.4.7}$$

then we know $p(R)$, which is known as a *probability density*. The proof follows by choosing

$$f(R) = \begin{cases} 1 & R_0 \leqslant R < R_0 + dR_0\,, \\ 0 & \text{otherwise}\,. \end{cases} \tag{2.4.8}$$

Because the expectation of an arbitrary function is sometimes a little easier to work with than a density, this relation will be used occasionally in this book.

Notation: The notation $\langle A \rangle$ for the expectation used in this book is a physicist's notation. The most common mathematical notation is $E(A)$, which is in my opinion a little less intuitive.

2.4.2 Sets of Probability Zero

If a density $p(R)$ exists, the probability that R is in the interval $(R_0, R_0 + dR)$ goes to zero with dR. Hence, the probability that R has *exactly* the value R_0 is zero; and similarly for any other value.

Thus, in such a case, there are sets $S\,(R_i)$, each containing only one point R_i, which have zero probability. From probability axiom iii), any countable union of such sets, i.e., any set containing only a countable number of points (e.g., all rational numbers) has probability zero. In general, all equalities in probability theory are at best only "*almost certainly true*", i.e., they may be untrue on sets of probability zero. Alternatively, one says, for example,

$$X = Y \text{ with probability } 1\,, \tag{2.4.9}$$

which is by no means the same as saying that

$$X(R) = Y(R) \text{ for all } R\,. \tag{2.4.10}$$

Of course, if the theory is to have any connection with reality, events with probability zero do not occur.

In particular, notice that our previous result if inspected carefully, only implies that we know $p(R)$ only with probability 1, given that we know $\langle f(R)\rangle$ for all $f(R)$.

2.5 The Interpretation of Mean Values

The question of what to measure in a probabilistic system is nontrivial. In practice, one measures either a set of individual values of a random variable (the number of animals of a certain kind in a certain region at certain points in time; the electric current passing through a given circuit element in each of a large number of replicas of that circuit, etc.) or alternatively, the measuring procedure may implicitly construct an average of some kind. For example, to measure an electric current, we may measure the electric charge transferred and divide by the time taken—this gives a measure of the average number of electrons transferred per unit time. It is important to note the essential difference in this case, that it will not normally be possible to measure anything other than a few selected averages and thus, higher moments (for example) will be unavailable.

In contrast, when we measure individual events (as in counting animals), we can then construct averages of the observables by the obvious method

$$\bar{X}_N = \frac{1}{N}\sum_{n=1}^{N} X(n)\,. \tag{2.5.1}$$

The quantities $X(n)$ are the individual observed values of the quantity X. We expect that as the number of samples N becomes very large, the quantity $\bar{X}_N$ approaches the mean $\langle X\rangle$ and that, in fact,

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N} f[X(n)] = \lim_{N\to\infty}\overline{f(X)}_N = \langle f(X)\rangle \tag{2.5.2}$$

and such a procedure will determine the probability density function $p(x)$ of X if we carry out this procedure for all functions f. The validity of this procedure depends on the degree of independence of the successive measurements and is dealt with in Sect. 2.5.2.

In the case where only averages themselves are directly determined by the measuring method, it will not normally be possible to measure $X(n)$ and therefore, it will not, in general, be possible to determine $\overline{f(X)}_N$. All that will be available will be $f(\bar{X}_N)$—quite a different thing unless f is linear. We can often find situations in which measurable quantities are related (by means of some theory) to mean values of certain functions, but to hope to measure, for example, the mean value of an arbitrary function of the number of electrons in a conductor is quite hopeless. The mean number—yes, and indeed even the mean square number, but the measuring methods available are not direct. We do *not* enumerate the individual numbers of electrons at different times and hence arbitrary functions are not attainable.

2.5.1 Moments, Correlations, and Covariances

Quantities of interest are given by the *moments* $\langle X^n \rangle$ since these are often easily calculated. However, probability densities must always vanish as $x \to \pm\infty$, so we see that higher moments tell us only about the properties of unlikely large values of X. In practice we find that the most important quantities are related to the first and second moments. In particular, for a single variable X, *the variance* defined by

$$\text{var}[X] \equiv \{\sigma[X]\}^2 \equiv \langle [X - \langle X \rangle]^2 \rangle , \tag{2.5.3}$$

and as is well known, the *variance* var$[X]$ or its square root the *standard deviation* $\sigma[X]$, is a measure of the degree to which the values of X deviate from the mean value $\langle X \rangle$.

In the case of several variables, we define the *covariance matrix* as

$$\langle X_i, X_j \rangle \equiv \langle (X_i - \langle X_i \rangle)(X_j - \langle X_j \rangle) \rangle \equiv \langle X_i X_j \rangle - \langle X_i \rangle \langle X_j \rangle . \tag{2.5.4}$$

Obviously,

$$\langle X_i, X_i \rangle = \text{var}[X_i]. \tag{2.5.5}$$

If the variables are independent *in pairs*, the covariance matrix is diagonal.

2.5.2 The Law of Large Numbers

As an application of the previous concepts, let us investigate the following model of measurement. We assume that we measure the same quantity N times, obtaining sample values of the random variable $X(n)$; $(n = 1, 2, \ldots, N)$. Since these are all measurements of the same quantity at successive times, we assume that for every $n, X(n)$ has the same probability distribution but we do not assume the $X(n)$ to be independent. However, provided the covariance matrix $\langle X(n), X(m) \rangle$ vanishes sufficiently rapidly as $|n - m| \to \infty$, then defining

$$\bar{X}_N = \frac{1}{N} \sum_{n=1}^{N} X(n) , \tag{2.5.6}$$

we shall show

$$\lim_{N\to\infty} \bar{X}_N = \langle X \rangle . \tag{2.5.7}$$

It is clear that

$$\langle \bar{X}_N \rangle = \langle X \rangle . \tag{2.5.8}$$

We now calculate the variance of $\bar{X}_N$ and show that as $N \to \infty$ it vanishes under certain conditions:

$$\langle \bar{X}_N \bar{X}_N \rangle - \langle \bar{X}_N \rangle^2 = \frac{1}{N^2} \sum_{n,m=1}^{N} \langle X_n, X_m \rangle \,. \tag{2.5.9}$$

Provided $\langle X_n, X_m \rangle$ falls off sufficiently rapidly as $|n - m| \to \infty$, we find

$$\lim_{N\to\infty} (\mathrm{var}[\bar{X}_N]) = 0 \,, \tag{2.5.10}$$

so that $\lim\limits_{N\to\infty} \bar{X}_N$ is a deterministic variable equal to $\langle X \rangle$.

Two models of $\langle X_n, X_m \rangle$ can be chosen.

a) $$\langle X_n, X_m \rangle \sim K\lambda^{|m-n|} \,, \qquad (\lambda < 1) \,, \tag{2.5.11}$$

for which one finds

$$\mathrm{var}[\bar{X}_N] = \frac{2K}{N^2}\left(\frac{\lambda^{N+2} - N(\lambda - 1) - \lambda}{(\lambda - 1)^2}\right) - \frac{K}{N} \to 0 \,. \tag{2.5.12}$$

b) $$\langle X_n, X_m \rangle \sim |n - m|^{-1} \,, \qquad (n \neq m) \,, \tag{2.5.13}$$

and one finds approximately

$$\mathrm{var}[\bar{X}_N] \sim \frac{2}{N} \log N - \frac{1}{N} \to 0 \,. \tag{2.5.14}$$

In both these cases, $\mathrm{var}[X_N] \to 0$, but the rate of convergence is very different. Interpreting n, m as the times at which the measurement is carried out, one sees than even very slowly decaying correlations are permissible. The law of large numbers comes in many forms, which are nicely summarised by *Papoulis* [2.2]. The central limit theorem is an even more precise result in which the limiting distribution function of $\bar{X}_N - \langle X \rangle$ is determined (see Sect. 2.8.2).

2.6 Characteristic Function

One would like a condition where the variables are independent, not just in pairs. To this end (and others) we define the characteristic function.

If $\boldsymbol{s}$ is the vector $(s_1, s_2, \ldots, s_n)$, and $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a vector of random variables, then the characteristic function (or moment generating function) is defined by

$$\phi(\boldsymbol{s}) = \langle \exp(\mathrm{i}\boldsymbol{s} \cdot \boldsymbol{X}) \rangle = \textstyle\int d\boldsymbol{x} \; p(\boldsymbol{x}) \exp(\mathrm{i}\boldsymbol{s} \cdot \boldsymbol{x}) \,. \tag{2.6.1}$$

The characteristic function has the following properties ([2.1], Chap. XV)

i) $\phi(\mathbf{0}) = 1$.

ii) $|\phi(\boldsymbol{s})| \leqslant 1$.

iii) $\phi(\boldsymbol{s})$ is a uniformly continuous function of its arguments for all finite real $\boldsymbol{s}$ [2.4].

iv) If the *moments* $\langle \prod\limits_i X_i^{m_i} \rangle$ exist, then

$$\left\langle \prod_i X_i^{m_i} \right\rangle = \left[\prod_i \left(-\mathrm{i} \frac{\partial}{\partial s_i} \right)^{m_i} \phi(\boldsymbol{s}) \right]_{\boldsymbol{s}=\mathbf{0}} \,. \tag{2.6.2}$$

v) A sequence of probability densities converges to limiting probability density if and only if the corresponding characteristic functions converge to the corresponding characteristic function of the limiting probability density.

vi) Fourier inversion formula

$$p(\boldsymbol{x}) = (2\pi)^{-n} \int d\boldsymbol{s}\, \phi(\boldsymbol{s}) \exp(-\mathrm{i}\boldsymbol{x} \cdot \boldsymbol{s})\,. \tag{2.6.3}$$

Because of this inversion formula, $\phi(\boldsymbol{s})$ determines $p(\boldsymbol{x})$ with probability 1. Hence, the characteristic function does truly *characterise* the probability density.

vii) Independent random variables: from the definition of independent random variables in Sect. 2.3.4, it follows that the variables $X_1, X_2 \dots$ are independent if and only if

$$p(x_1, x_2, \dots, x_n) = p_1(x_1) p_2(x_2) \dots p_n(x_n)\,, \tag{2.6.4}$$

in which case,

$$\phi(s_1, s_2, \dots s_n) = \phi_1(s_1)\phi_2(s_2) \dots \phi_n(s_n). \tag{2.6.5}$$

viii) Sum of independent random variables: if $X_1, X_2, \dots,$ are independent random variables and if

$$Y = \sum_{i=1}^{n} X_i\,, \tag{2.6.6}$$

and the characteristic function of Y is

$$\phi_y(s) = \langle \exp(\mathrm{i}sY) \rangle\,, \tag{2.6.7}$$

then

$$\phi_y(s) = \prod_{i=1}^{n} \phi_i(s)\,. \tag{2.6.8}$$

The characteristic function plays an important role in this book which arises from the convergence property (v), which allows us to perform limiting processes on the characteristic function rather than the probability distribution itself, and often makes proofs easier. Further, the fact that the characteristic function is truly characteristic, i.e., the inversion formula (vi), shows that different characteristic functions arise from different distributions. As well as this, the straightforward derivation of the moments by (2.6.2) makes any determination of the characteristic function directly relevant to measurable quantities.

2.7 Cumulant Generating Function: Correlation Functions and Cumulants

A further important property of the characteristic function arises by considering its logarithm

$$\Phi(\boldsymbol{s}) = \log \phi(\boldsymbol{s})\,, \tag{2.7.1}$$

which is called the *cumulant generating function*. Let us assume that all moments exist so that $\phi(\boldsymbol{s})$ and hence, $\Phi(\boldsymbol{s})$, is expandable in a power series which can be written as

$$\Phi(\boldsymbol{s}) = \sum_{r=1}^{\infty} \mathrm{i}^r \sum_{\{m\}} \langle\langle X_1^{m_1} X_2^{m_2} \dots X_n^{m_n} \rangle\rangle \frac{s_1^{m_1} s_2^{m_2} \dots s_n^{m_n}}{m_1! m_2! \dots m_n!} \, \delta\left(r, \sum_{i=1}^{n} m_i\right), \tag{2.7.2}$$

where the quantities $\langle\langle X_1^{m_1} X_2^{m2} \dots X_n^{m_n} \rangle\rangle$ are called the *cumulants* of the variables $\boldsymbol{X}$. The notation chosen should not be taken to mean that the cumulants are functions of the particular product of powers of the $\boldsymbol{X}$; it rather indicates the moment of highest order which occurs in their expression in terms of moments. *Stratonovich* [2.5] also uses the term *correlation functions*, a term which we shall reserve for cumulants which involve more than one X_i. For, if the $\boldsymbol{X}$ are all independent, the factorisation property (2.6.6) implies that $\Phi(\boldsymbol{s})$ (the cumulant generating function) is a sum of n terms, each of which is a function of only one s_i and hence the coefficient of mixed terms, i.e., the *correlation functions* (in our terminology) are all zero and the converse is also true. Thus, the magnitude of the correlation functions is a measure of the degree of correlation.

The cumulants and correlation functions can be evaluated in terms of moments by expanding the characteristic function as a power series:

$$\phi(\boldsymbol{s}) = \sum_{r=1}^{\infty} \frac{\mathrm{i}^r}{r!} \sum_{\{m\}} \langle X_1^{m_1} X_2^{m_2} \dots X_n^{m_n} \rangle \frac{r!}{m_1! m_2! \dots m_n!} \, \delta\left(r, \sum_{i=1}^{n} m_i\right) s_1^{m_1} s_2^{m_2} \dots s_n^{m_n} \,. \tag{2.7.3}$$

Expanding the logarithm in a power series, and comparing it with (2.7.2) for $\Phi(\boldsymbol{s})$, the relationship between the cumulants and the moments can be deduced. No *simple* formula can be given, but the first few cumulants can be exhibited: we find

$$\langle\langle X_i \rangle\rangle = \langle X_i \rangle \,, \tag{2.7.4}$$

$$\langle\langle X_i X_j \rangle\rangle = \langle X_i X_j \rangle - \langle X_i \rangle \langle X_j \rangle \,, \tag{2.7.5}$$

$$\langle\langle X_i X_j X_k \rangle\rangle = \langle X_i X_j X_k \rangle - \langle X_i X_j \rangle \langle X_k \rangle - \langle X_i \rangle \langle X_j X_k \rangle - \langle X_i X_k \rangle \langle X_j \rangle + 2 \langle X_i \rangle \langle X_j \rangle \langle X_k \rangle \,. \tag{2.7.6}$$

Here, all formulae are also valid for any number of equal i, j, k, l. An explicit general formula can be given as follows. Suppose we wish to calculate the cumulant $\langle\langle X_1 X_2 X_3 \dots X_n \rangle\rangle$. The procedure is the following:

i) Write a sequence of n dots ...... ;

ii) Divide into $p + 1$ subsets by inserting angle brackets

$$\langle \dots \rangle \langle . . \rangle \langle \dots\dots \rangle . . \langle . . \rangle \,; \tag{2.7.7}$$

iii) Distribute the symbols $X_1 \dots X_n$ in place of the dots in such a way that all *different* expressions of this kind occur, e.g.,

$$\langle X_1 \rangle \langle X_2 X_3 \rangle = \langle X_1 \rangle \langle X_3 X_2 \rangle \neq \langle X_3 \rangle \langle X_1 X_2 \rangle \,; \tag{2.7.8}$$

iv) Take the sum of all such terms for a given p. Call this $C_p(X_1, X_2, \dots, X_n)$;

v) $$\langle\langle X_1 X_2 \dots X_n \rangle\rangle = \sum_{p=0}^{n-1} (-1)^p p! C_p(X_1, X_2, \dots, X_n)\,. \tag{2.7.9}$$

A derivation of this formula was given by *Meeron* [2.6]. The particular procedure is due to *van Kampen* [2.7].

vi) *Cumulants in which there is one or more repeated element*:
For example $\langle\langle X_1^2 X_3 X_2 \rangle\rangle$—simply evaluate $\langle\langle X_1 X_2 X_3 X_4 \rangle\rangle$ and set $X_4 = X_1$ in the resulting expression.

2.7.1 Example: Cumulant of Order 4: $\langle\langle X_1 X_2 X_3 X_4 \rangle\rangle$

a) $p = 0$

Only term is $\langle X_1 X_2 X_3 X_4 \rangle = C_0(X_1 X_2 X_3 X_4)$.

b) $p = 1$

Partition $\langle . \rangle\langle \dots \rangle$
Term $\{\langle X_1 \rangle\langle X_2 X_3 X_4 \rangle + \langle X_2 \rangle\langle X_3 X_4 X_1 \rangle + \langle X_3 \rangle\langle X_4 X_1 X_2 \rangle$
$+\langle X_4 \rangle\langle X_1 X_2 X_3 \rangle\} \equiv D_1$

partition $\langle . . \rangle\langle . . \rangle$
Term $\langle X_1 X_2 \rangle\langle X_3 X_4 \rangle + \langle X_1 X_3 \rangle\langle X_2 X_4 \rangle + \langle X_1 X_4 \rangle\langle X_2 X_3 \rangle \equiv D_2\,.$

Hence,

$$D_1 + D_2 = C_1(X_1 X_2 X_3 X_4)\,. \tag{2.7.10}$$

c) $p = 2$

Partition $\langle . \rangle\langle . \rangle\langle . . \rangle$
Term $\langle X_1 \rangle\langle X_2 \rangle\langle X_3 X_4 \rangle + \langle X_1 \rangle\langle X_3 \rangle\langle X_2 X_4 \rangle + \langle X_1 \rangle\langle X_4 \rangle\langle X_2 X_3 \rangle$
$+\langle X_2 \rangle\langle X_3 \rangle\langle X_1 X_4 \rangle + \langle X_2 \rangle\langle X_4 \rangle\langle X_1 X_3 \rangle + \langle X_3 \rangle\langle X_4 \rangle\langle X_1 X_2 \rangle$
$= C_2(X_1 X_2 X_3 X_4)$.

d) $p = 3$

Partition $\langle . \rangle\langle . \rangle\langle . \rangle\langle . \rangle$
Term $\langle X_1 \rangle\langle X_2 \rangle\langle X_3 \rangle\langle X_4 \rangle = C_3(X_1 X_2 X_3 X_4)\,.$

Hence,

$$\langle\langle X_1 X_2 X_3 X_4 \rangle\rangle = C_0 - C_1 + 2C_2 - 6C_3\,. \tag{2.7.11}$$

2.7.2 Significance of Cumulants

From (2.7.4, 2.7.5) we see that the first two cumulants are the means $\langle X_i \rangle$ and covariances $\langle X_i, X_j \rangle$. Higher-order cumulants contain information of decreasing significance, unlike higher-order moments. We cannot set all *moments* higher than a certain order equal to zero since $\langle X^{2n} \rangle \geqslant \langle X^n \rangle^2$ and thus, all moments contain information about lower moments.

For cumulants, however, we can consistently set

$$\begin{aligned}\langle\!\langle X\rangle\!\rangle &= a\,,\\ \langle\!\langle X^2\rangle\!\rangle &= \sigma^2\,,\\ \langle\!\langle X^n\rangle\!\rangle &= 0\,, \qquad (n>2)\,,\end{aligned}$$

and we can easily deduce by using the inversion formula for the characteristic function that

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-a)^2}{2\sigma^2}\right), \tag{2.7.12}$$

that is, a Gaussian probability distribution. It does not, however, seem possible to give more than this intuitive justification. Indeed, the theorem of *Marcinkiewicz* [2.8, 2.9] shows that the cumulant generating function cannot be a polynomial of degree greater than 2, that is, either all but the first 2 cumulants vanish or there are an infinite number of nonvanishing cumulants. The greatest significance of cumulants lies in the definition of the correlation functions of different variables in terms of them; this leads further to important approximation methods.

2.8 Gaussian and Poissonian Probability Distributions

2.8.1 The Gaussian Distribution

By far the most important probability distribution is the Gaussian, or normal distribution. Here we collect together the most important facts about it.

If $\boldsymbol{X}$ is a vector of n Gaussian random variables, the corresponding multivariate probability density function can be written

$$p(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n\det(\sigma)}}\exp\left[-\tfrac{1}{2}(\boldsymbol{x}-\bar{\boldsymbol{x}})^{\mathrm{T}}\sigma^{-1}(\boldsymbol{x}-\bar{\boldsymbol{x}})\right], \tag{2.8.1}$$

so that

$$\langle\boldsymbol{X}\rangle = \textstyle\int d\boldsymbol{x}\,\boldsymbol{x}\,p(\boldsymbol{x}) = \bar{\boldsymbol{x}}, \tag{2.8.2}$$

$$\langle\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\rangle = \textstyle\int d\boldsymbol{x}\,\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}p(\boldsymbol{x}) = \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathrm{T}} + \sigma, \tag{2.8.3}$$

and the characteristic function is given by

$$\phi(\boldsymbol{s}) = \langle\exp(\mathrm{i}\boldsymbol{s}^{\mathrm{T}}\boldsymbol{X})\rangle = \exp\left(\mathrm{i}\boldsymbol{s}^{\mathrm{T}}\bar{\boldsymbol{x}} - \tfrac{1}{2}\boldsymbol{s}^{\mathrm{T}}\sigma\boldsymbol{s}\right). \tag{2.8.4}$$

This particularly simple characteristic function implies that all cumulants of higher order than 2 vanish, and hence means that all moments of order higher than 2 are expressible in terms of those of order 1 and 2. The relationship (2.8.3) means that σ is the covariance matrix (as defined in Sect. 2.5.1), i.e., the matrix whose elements are the second-order correlation functions. Of course, σ is symmetric.

The precise relationship between the higher moments and the covariance matrix σ can be written down straightforwardly by using the relationship between the moments and the characteristic function [Sect. 2.6 iv)]. The formula is only simple if $\bar{\boldsymbol{x}} = 0$, in which case the odd moments vanish and the even moments satisfy

$$\langle X_i X_j X_k \dots \rangle = \frac{(2N)!}{N!2^N} \{\sigma_{ij}\sigma_{kl}\sigma_{mn} \dots\}_{\text{sym}} \,, \tag{2.8.5}$$

where the subscript "sym" means the symmetrised form of the product of σ's, and $2N$ is the order of the moment. For example,

$$\begin{aligned}\langle X_1 X_2 X_3 X_4 \rangle &= \frac{4!}{4.2!}\left\{\frac{1}{3}[\sigma_{12}\sigma_{34} + \sigma_{41}\sigma_{23} + \sigma_{13}\sigma_{24}]\right\} \,, \\ &= \sigma_{12}\sigma_{34} + \sigma_{41}\sigma_{23} + \sigma_{13}\sigma_{24} \,, \end{aligned} \tag{2.8.6}$$

$$\langle X_1^4 \rangle = \frac{4!}{4.2!}\left\{\sigma_{11}^2\right\} = 3\sigma_{11}^2 \,. \tag{2.8.7}$$

2.8.2 Central Limit Theorem

The Gaussian distribution is important for a variety of reasons. Many variables are, in practice, empirically well approximated by Gaussians and the reason for this arises from the *central limit theorem*, which, roughly speaking, asserts that a random variable composed of the sum of many parts, each independent but arbitrarily distributed, is Gaussian. More precisely, let $X_1, X_2, X_3, \dots, X_n$ be independent random variables such that

$$\langle X_i \rangle = 0, \qquad \text{var}[X_i] = b_i^2, \tag{2.8.8}$$

and let the distribution function of X_i be $p_i(x_i)$.

Define

$$\mathrm{S}_n = \sum_{i=1}^{n} X_i \,, \tag{2.8.9}$$

and

$$\sigma_n^2 = \text{var}[S_n] = \sum_{i=1}^{n} b_i^2 \,. \tag{2.8.10}$$

We require further the fulfilment of the *Lindeberg condition*:

$$\lim_{n\to\infty}\left[\frac{1}{\sigma_n^2}\sum_{i=1}^{n}\int_{|x|>t\sigma_n} dx\, x^2\, p_i(x)\right] = 0, \tag{2.8.11}$$

for any fixed $t > 0$. Then, under these conditions, the distribution of the normalised sums S_n/σ_n tends to the Gaussian with zero mean and unit variance.

The proof of the theorem can be found in [2.1]. It is worthwhile commenting on the hypotheses, however. We first note that the summands X_i are required to be independent. This condition is not absolutely necessary; for example, choose

$$X_i = \sum_{r=i}^{i+j} Y_r, \tag{2.8.12}$$

where the Y_j are independent. Since the sum of the X's can be rewritten as a sum of Y's (with certain finite coefficients), the theorem is still true.

Roughly speaking, as long as the correlation between X_i and X_j goes to zero sufficiently rapidly as $|i-j| \to \infty$, a central limit theorem will be expected. The Lindeberg condition (2.8.11) is not an obviously understandable condition but is the weakest condition which expresses the requirement that the probability for $|X_i|$ to be large is very small. For example, if all the b_i are infinite or greater than some constant C, it is clear that σ_n^2 diverges as $n \to \infty$. The sum of integrals in (2.8.11) is the sum of contributions to variances for all $|X_i| > t\sigma_n$, and it is clear that as $n \to \infty$, each contribution goes to zero. The Lindeberg condition requires the sum of all the contributions not to diverge as fast as σ_n^2. In practice, it is a rather weak requirement; satisfied if $|X_i| < C$ for all X_i, or if $p_i(x)$ go to zero sufficiently rapidly as $x \to \pm\infty$. An exception is

$$p_i(x) = \frac{a_i}{\pi(x^2 + a_i^2)}, \tag{2.8.13}$$

the *Cauchy*, or *Lorentzian* distribution. The variance of this distribution is infinite and, in fact, the sum of all the X_i has a distribution of the same form as (2.8.13) with a_i replaced by $\sum\limits_{i=1}^{n} a_i$. Obviously, the Lindeberg condition is not satisfied.

A related condition, also called the Lindeberg condition, will arise in Sect. 3.3.1, where we discuss the replacement of a discrete process by one with continuous steps.

2.8.3 The Poisson Distribution

A distribution which plays a central role in the study of random variables which take on positive integer values is the Poisson distribution. If X is the relevant variable the Poisson distribution is defined by

$$P(X = x) \equiv P(x) = \frac{\mathrm{e}^{-\alpha}\alpha^x}{x!}, \tag{2.8.14}$$

and clearly, the *factorial moments*, defined by

$$\langle X^r \rangle_{\mathrm{f}} = \langle x(x-1)\ldots(x-r+1)\rangle, \tag{2.8.15}$$

are given by

$$\langle X^r \rangle_{\mathrm{f}} = \alpha^r. \tag{2.8.16}$$

For variables whose range is nonnegative integral, we can very naturally define the *generating function*

$$G(s) = \sum_{x=0}^{\infty} s^x P(x) = \langle s^x \rangle, \tag{2.8.17}$$

which is related to the characteristic function by

$$G(s) = \phi(-\mathrm{i}\log s). \tag{2.8.18}$$

The generating function has the useful property that

$$\langle X^r \rangle_{\mathrm{f}} = \left[\left(\frac{\partial}{\partial s}\right)^r G(s)\right]_{s=1}. \tag{2.8.19}$$

For the Poisson distribution we have

$$G(s) = \sum_{x=0}^{\infty} \frac{e^{-\alpha}(s\alpha)^x}{x!} = \exp[\alpha(s-1)]\,. \tag{2.8.20}$$

We may also define the factorial cumulant generating function $g(s)$ by

$$g(s) = \log G(s) \tag{2.8.21}$$

and the *factorial cumulants* $\langle\langle X^r \rangle\rangle_f$ by

$$g(s) = \sum_{x=1}^{\infty} \langle\langle X^r \rangle\rangle_f \frac{(s-1)^r}{r!}\,. \tag{2.8.22}$$

We see that the Poisson distribution has all but the first factorial cumulant zero.

The Poisson distribution arises naturally in very many contexts, for example, we have already met it in Sect. 1.5.1 as the solution of a simple master equation. It plays a similar central role in the study of random variables which take on integer values to that occupied by the Gaussian distribution in the study of variables with a continuous range. However, the only simple multivariate generalisation of the Poisson is simply a product of Poissons, i.e., of the form

$$P(x_1, x_2, x_3, \dots) = \prod_{i=1}^{n} \frac{e^{-\alpha_i}(\alpha_i)^{x_i}}{x_i!}\,. \tag{2.8.23}$$

There is no logical concept of a correlated multipoissonian distribution, similar to that of a correlated multivariate Gaussian distribution.

2.9 Limits of Sequences of Random Variables

Much of computational work consists of determining *approximations* to random variables, in which the concept of a *limit of a sequence of random variables* naturally arises. However, there is no unique way of defining such a limit.

For, suppose we have a probability space Ω, and a sequence of random variables X_n defined on Ω. Then by the limit of the sequence as $n \to \infty$

$$X = \lim_{n\to\infty} X_n\,, \tag{2.9.1}$$

we mean a random variable X which, in some sense, is approached by the sequence of random variables X_n. The various possibilities arise when one considers that the probability space Ω has elements ω which have a probability density $p(\omega)$. Then we can choose the following definitions.

2.9.1 Almost Certain Limit

X_n converges *almost certainly* to X if, for all ω except a set of probability zero

$$\lim_{n\to\infty} X_n(\omega) = X(\omega)\,. \tag{2.9.2}$$

Thus each realisation of X_n converges to X and we write

$$\operatorname*{ac\text{-}lim}_{n\to\infty} X_n = X. \tag{2.9.3}$$

2.9.2 Mean Square Limit (Limit in the Mean)

Another possibility is to regard the $X_n(\omega)$ as functions of ω, and look for the mean square deviation of $X_n(\omega)$ from $X(\omega)$. Thus, we say that X_n converges to X in the *mean square* if

$$\lim_{n\to\infty} \int d\omega\, p(\omega)[X_n(\omega) - X(\omega)]^2 \equiv \lim_{n\to\infty} \langle (X_n - X)^2 \rangle = 0. \tag{2.9.4}$$

This is the kind of limit which is well known in Hilbert space theory. We write

$$\operatorname*{ms-lim}_{n\to\infty} X_n = X\,. \tag{2.9.5}$$

2.9.3 Stochastic Limit, or Limit in Probability

We can consider the possibility that $X_n(\omega)$ approaches X because the probability of deviation from X approaches zero: precisely, this means that if for any $\varepsilon > 0$

$$\lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0\,, \tag{2.9.6}$$

then the *stochastic limit* of X_n is X.

In this case, we write

$$\operatorname*{st-lim}_{n\to\infty} X_n = X\,. \tag{2.9.7}$$

2.9.4 Limit in Distribution

An even weaker form of convergence occurs if, for any continuous bounded function $f(x)$

$$\lim_{n\to\infty} \langle f(X_n) \rangle = \langle f(X) \rangle\,. \tag{2.9.8}$$

In this case the convergence of the limit is said to be *in distribution*. In particular, using $\exp(\mathrm{i}xs)$ for $f(x)$, we find that the characteristic functions approach each other, and hence the probability density of X_n approaches that of X.

2.9.5 Relationship Between Limits

The following relations can be shown.

1) Almost certain convergence $\Longrightarrow$ stochastic convergence.

2) Convergence in mean square $\Longrightarrow$ stochastic convergence.

3) Stochastic convergence $\Longrightarrow$ convergence in distribution.

All of these limits have uses in applications.