Part I

Basic principles

1

What is a stochastic process?

Intuitively, a stochastic process describes some phenomenon that evolves over time (a *process*) and that involves a random (a *stochastic*) component. Empirically, we observe such a process by recording values of an appropriate response variable at various points in time. In interesting cases, the phenomenon under study will usually depend on covariates. Some of these may also vary over time, whereas others will define the differing (static) conditions under which various 'copies' of the process occur.

1.1 Definition

Thus, a stochastic process involves some response variable, say Y_t, that takes values varying randomly in some way over time t (or space, although that will not be considered here). Y_t may be a scalar or a vector, but I shall concentrate primarily on scalar responses in this text (see, however, Section 8.3). Generally, in the study of such processes, the term 'random' is replaced by '*stochastic*'; hence, the name.

An observed value or realisation y_t of the response variable Y_t is called the *state* of the process at time t. We might call an observation of the state of a process an event. However, I shall restrict the meaning of *event* to the occurrence of a *change* of state. Thus, the number of possible different events will depend, among other things, on the number of distinct states.

More generally, the probability of the process being in some given state at some point in time may depend on some function of previous events and of covariates. Usually, the probabilities of possible events will be conditional on the state of the process. Such relationships will thus be determined by the type of model being fitted.

The main properties distinguishing among observed stochastic processes are:

(i) The frequency or periodicity with which observations are made over time.
(ii) The set of all of its possible observable values, that is, of all possible responses or states of the series, called the *state space*.
(iii) The sources and forms of randomness present, including the nature of the dependence among the values in a series of realisations of the random variable Y_t.

(iv) The number of 'copies' of the process available (only one or several), which will determine how adequate information can be obtained for modelling.

Let us look more closely at each of these aspects.

1.1.1 Time

Observations of a stochastic process (at least of the kinds that interest us here) are made over *time*. If these observations are at equally spaced intervals, time is said to be *discrete*. Otherwise, time is *continuous*. Notice, however, that a process can never really be observed continuously because that would imply an infinite number of observations, even in a small interval of time. Hence, the distinction is primarily used to determine what kind of model will be applied. Continuous-time models can be used for any data but may be more complex, and may not be appropriate if changes can only occur at discrete time points. Discrete-time models require equally-spaced observations, unless some simple mechanism for missingness can be introduced.

Attention may centre on

(i) the *states* at given points in time,
(ii) the *events*, that is, on what change of state occurs at each particular time point, or
(iii) the *times* of events.

Consider simple examples:

- When an economist measures monthly unemployment, time is only an equally-spaced indicator of when observations are made. The number of unemployed may change at each monthly recording. Either the level of employment (the state) or the amount of change (the event) may be of central interest. Time itself is not of direct concern except for ordering the observations.
- In contrast, when a doctor records the times of a patient's repeated infections, the state might be defined to be the total number of such infections so far suffered by that patient. Each observation (event) is the same, an infection, and the time between these events is essential. With substantial loss of information, this could be reduced to discrete time by recording only the numbers of infections in equally-spaced time intervals.

A response may only be recorded when some specific event of interest occurs. However, in order to determine the timing of that event, fairly continual observation of the process usually is necessary, that is, a series of intermediate, implicit recordings of no event. If observation begins at some natural time point, such as birth, at which no response value occurs, the mechanism determining the time to the first event will usually be different from that between subsequent events.

1.1.2 State space

At any given point in time, a process will be in some *state*. This usually is observed by recording the value of a response variable y_t at that point t. As always in statistical modelling, the set of possible states is a scientific construct that should be defined in a way appropriate to answer the questions at hand. Although, in principle, the state is what is being observed, certain models also assume a second process of unobservable *hidden* states (Chapters 7 and 11).

The set of all possible observable states is called the *state space*. This may be finite, yielding a categorical response variable, or infinite, giving either a discrete or a continuous response variable. Generally, a *minimal* state space is chosen, one in which the probability (density) of every state is nonzero.

If an observed response variable were truly continuous, every recording would be an event because no two would be the same. However, this is empirically impossible so that any observable process could possibly stay in the same state (given the limit of precision of the recording instrument) over two or more consecutive observation points.

A categorical response usually refers to a finite set of possible different states that may be observed. However, when only one type of response is of particular interest and it is fairly rare, the response might be recorded as binary, indicating presence (coded 1) or absence (coded 0) of that response. This particular response, with a binary state space, is often called a *point event* or a *recurrent event* (Chapter 4); thus, here the term 'event' refers both to one of the states and to the change of state. Above, I gave an example of repeated infections, where the cumulative number of infections was important. If, instead, repeated epileptic fits were recorded, the states might more appropriately be defined as having a fit or not on each particular day instead of the total number of fits so far.

In certain situations, the state may be defined by more than one response value, that is, $\mathbf{y}_t$ may be a vector containing quite distinct types of values. Thus, there may be a categorical value, such as a binary indicator that it rains or not, accompanied by a (usually quantitative) value, the *mark*, for example, how much rain fell (Section 8.3).

A vector $\mathbf{y}_t$ also is usually necessary when there are *endogenous* time-varying covariates. These are variables, other than the response of direct interest, that are influenced by the previous states of that response. Suppose, for example, that the condition of a patient, measured in some appropriate way, is the response of direct interest. If the dose of a medication that a patient receives depends upon his or her previous condition, then dose and condition will generally have to be handled simultaneously. They must together define the state and be allowed to vary stochastically interdependently, if a reasonable model is to be constructed.

A process may also involve time-varying *exogenous* covariates, that is, variables not influenced by the previous states of the process. The stochastic variability of such covariates is not usually of interest, so that the probability of the state y_t can be taken to be conditional on their observed values, as in classical regression models.

Practically, we see from this section and the preceding one that models for

Table 1.1. *Chapters in which different types of stochastic processes are covered.*

	State space	
Time	Categorical	Continuous
Discrete	5, 8	9
Continuous	3, 4, 6, 7, 8	9, 10, 11, 12, 13, 14

stochastic processes can be classified by the type of variable observed, that is, the state space, and by the frequency or regularity with which it is observed over time. The structure of this text according to these criteria is summarised in Table 1.1. Recall, however, that most models in continuous time also can be applied to observations in discrete time.

1.1.3 Randomness

In a *deterministic* process, one can predict the exact sequence of outcomes (states) from the initial conditions, although some, especially chaotic, systems are extremely sensitive to these conditions. In contrast, a *stochastic* process has an inherent component of unpredictability or randomness. In empirical observation, randomness is closely associated with unknownness or incomplete information. To handle such situations, the theory of probability is used. Thus, predictions involving stochastic processes will not indicate specific outcomes, but only the probabilities of occurrence of the different possible outcomes.

A stochastic process may involve many forms of randomness. A number will arise from the imperfections of the observation and modelling procedures. However, in scientific modelling, the most important should be inherent in the process under study.

The types of randomness that need to be allowed for in modelling a stochastic process can have several sources including the following:

(i) Unmeasurable variability is built into many scientific systems. This is true of quantum mechanics, but also of most biological and social processes.
(ii) Almost invariably, all of the measurable factors determining the process cannot be take into account. For example, unrecorded environmental conditions may change over the period in which the process is acting.
(iii) The initial conditions may be very difficult to determine precisely.

Unfortunately, an inappropriate model can also generate additional spurious randomness and dependencies.

Traditionally, essentially for mathematical simplicity, the Poisson, binomial, and normal distributions have been used in modelling the randomness of stochastic processes. However, we shall see as we proceed that a wide variety of other distributions may be more suitable in specific circumstances.

1.1.4 Stationarity, equilibrium, and ergodicity

The series of responses y_t of a stochastic process usually will not be independent. Thus, procedures must be available to introduce appropriate dependencies. Because complex models should generally be avoided in science, relatively simple methods for introducing such dependencies are desirable.

Multivariate distributions

Multivariate probability distributions provide the general framework in which to specify the ways in which responses are interdependent. Unfortunately, in the context of stochastic processes, these may be difficult to use, especially if long series of observations are available. Thus, with a series of R observations, a model involving a multivariate normal distribution will require the manipulation of an $R \times R$ covariance matrix. When R is large, this will often not be practical or efficient, at least for technical reasons.

Fortunately, for an ordered series of R responses, such as those that interest us, the multivariate distribution always can be decomposed into an ordered sequence of independent *conditional* univariate distributions:

$$f(y_0, y_1, \ldots, y_R) = f_0(y_0) f_1(y_1|y_0) \cdots f_R(y_R|y_0, \ldots, y_{R-1}) \tag{1.1}$$

where

$$f_t(y_t|y_0, \ldots, y_{t-1}) = \frac{f(y_0, \ldots, y_t)}{f(y_0, \ldots, y_{t-1})} \tag{1.2}$$

Notice that each conditional distribution $f_t(y_t|y_0, \ldots, y_{t-1})$ may be completely different from all the others, even though the multivariate distributions $f(\cdot)$ for different lengths of series will usually have the same form.

Generally, it will be easier to work with the conditional distributions than the multivariate one. For the multivariate normal distribution, this will be a series of univariate normal distributions, not involving directly the covariance matrix. I shall elaborate on this approach in Section 1.2 below.

In contrast to the multivariate distribution, and its univariate conditional decomposition, the series of univariate *marginal* distributions, although often of interest, cannot by itself specify a unique stochastic process (unless all successive states are independent). It is generally of limited direct use in constructing a realistic model and will, most often, rather be a byproduct of the construction. On the other hand, this sequence of univariate marginal distributions of a stochastic process does provide valuable information indicating how the process is evolving over time: its 'underlying' profile or trend.

In the decomposition in Equation (1.1), no restrictive assumptions have been made (except ordering). However, in order for such a multivariate model to be tractable, even fitted conditionally, rather strong assumptions often have to be made. The problem with the general specification just given is that every response has a different conditional distribution, and each new response will have yet another one. The situation is changing faster than information can be accumulated about it! Thus, we require some reasonable simplifying assumptions.

Stationarity

A stochastic process is said to be *strictly stationary* if all sequences of consecutive responses of equal length in time have identical multivariate distributions

$$f(Y_1 = y_1, \cdots, Y_R = y_R) = f(Y_{t+1} = y_1, \cdots, Y_{t+R} = y_R) \tag{1.3}$$

for all t and R. In other words, shifting a fixed-width time observation window along a strictly stationary series always yields the same multivariate distribution. Such an assumption reduces enormously the amount of empirical information necessary in order to model a stochastic process.

A less restrictive assumption, reducing even further the amount of information required, is that a process is *second-order stationary*. This is defined only by the mean, variance, and covariances:

$$\begin{aligned} \mathrm{E}(Y_t) &= \mu \\ \mathrm{E}[(Y_t - \mu)(Y_{t+h} - \mu)] &= \gamma(h) \end{aligned} \tag{1.4}$$

for all t and h. Because a multivariate normal distribution is completely defined by its first two moments, if the process is normal and second-order stationary, it is strictly stationary. This is not generally true of other distributions. In this text, stationarity will always be strict.

Stationarity is a characteristic of multivariate distributions. Thus, from Equation (1.1), it cannot be determined solely by the conditional distributions, but requires also that the initial marginal distribution $f_0(y_0)$ be specified. Of course, this, in turn, implies that the univariate marginal distributions at all other time points will also be known.

Stationarity can be an appropriate assumption if the stochastic process has no inherent time origin. However, in experimental situations, for example, where treatments are applied at a specific time point, this will not be true. The need for greater information due to lack of stationarity can often be compensated by studying replications of the process (Section 1.1.5).

Equilibrium

Although a process may not be stationary when it starts, it may reach an *equilibrium* after a sufficiently long time, independent of the initial conditions. In other words, if an equilibrium has been reached, the probability that the process is in each given state, or the proportion of time spent in each state, has converged to a constant that does not depend on the initial conditions. This generally implies that eventually the process approaches closely to a stationary situation in the sense that, if it initially had the equilibrium distribution of states, it would be stationary. (See Cox and Miller, 1965, pp. 9, 272.)

Ergodicity

The concept of *ergodicity* is closely related to that of equilibrium, although the former has various meanings in the literature on stochastic processes. Ergodic theorems provide identities between probability averages, such as an expected value,

and long-run averages over a single realisation of the process. Thus, if the equilibrium probability of being in a given state equals the proportion of a long time period spent in that state, this is called an ergodic property of the process. In a similar way, the law of large numbers can be generalised to stochastic processes. (See Cox and Miller, 1965, p. 292.)

Regeneration points

Another important concept for some stochastic processes is that of a *regeneration point*. This is a time instant at which the process returns to a specific state such that future evolution of the process does not depend on how that state was reached. In other words, whenever such a process arrives at a regeneration point, all of its previous history is forgotten.

A well known case is the renewal process (Section 4.1.4) describing times between recurrent events which, as its name suggests, starts over again at each such event. (See Daley and Vere-Jones, 1988, p. 13.)

I shall look at further general procedures for simplifying models of stochastic processes in Section 1.2 below.

1.1.5 Replications

When studying a stochastic process, two approaches to obtaining adequate information can be envisaged. One can either observe

(i) one series for a long enough period, if it is reasonably stable, or
(ii) several short 'replications' of the process, if they are reasonably similar.

In certain situations, one has no choice but to use replications. Thus, for example, with survival data (Chapter 3), one single event, say death, terminates the process so that the only way to proceed is by collecting information on a large number of individuals and by assuming that the process is identical for all of them.

Both approaches can create problems. The phenomenon under study may not be stable enough to be observed over a very long time, say due to problems of lack of stationarity as discussed above. It may only be possible to assume that shorter segments of a series are from the same stochastic process. On the other hand, with replications, one must be able to assume that the different series recorded do, in fact, represent the same stochastic process. In certain situations, such as survival data, it may not even be possible to check this strong assumption empirically.

When replications of a stochastic process are modelled, extreme care must be taken with the time scale. If it is not chronological time, problems may arise. For example, in experiments with biological organisms, time may be measured either from birth or from start of a treatment. If all births do not occur simultaneously and treatment is not started at the same time for all subjects, events that occur at similar times after beginning treatment may occur neither closely together chronologically nor at similar ages. This can create difficult problems of interpretation due to confounding.

In the examples that I shall analyse in this text, I shall use either one long series or a set of several short ones, depending both on the type of problem and on the kind of data available. Generally, in the second case, when replications are present, I shall assume, for simplicity, that they come from the same process, perhaps with any differences explainable by time-constant (interprocess) covariates. Only in Section 7.3 and in Chapter 14 shall I look at some standard ways of modelling the differences among a set of series not described by observed covariates.

1.2 Dependence among states

As we have seen, dependencies among successive responses of a stochastic process can be modelled by multivariate distributions or equivalently by the corresponding product of conditional distributions. Certain general procedures are available that I shall review briefly here. Some of them arise from time series analysis (Chapter 9) but are of much wider applicability.

1.2.1 Constructing multivariate distributions

In a model for a stochastic process, some specific stochastic mechanism is assumed to generate the states. We often can expect that states of a series observed more closely together in time will be more similar, that is, more closely related. In other words, the state at a given time point will generally be related to those recently produced: the probability of a given state will be conditional, in some way, on the values of the process previously generated.

In certain situations, an adequate model without such dependence may be constructed. It usually will require the availability of appropriate time-varying covariates. If these have been recorded, and perhaps an appropriate time trend specified, the present state, conditional on the covariates, should be independent of previous states. However, in many cases, this will not be possible, because of lack of information, or will not be desirable, perhaps because of the complexity of the model required or its lack of generality.

In order to model time dependencies among the successive states of a stochastic process of interest, we may choose a given form either for the conditional distributions, on the right-hand side of Equation (1.1), or for the multivariate distribution, on the left-hand side. In general, the conditional distribution will be different from the multivariate and marginal distributions, because the ratio of two multivariate distributions does not yield a conditional distribution of the same form except in very special circumstances such as the normal distribution. The one or the other will most often be intractable.

Because a limited number of useful non-normal multivariate distributions is available, suitable models often can only be obtained by direct construction of the conditional distribution. Thus, usually, we shall need to set up some hierarchical series of conditional distributions, as in Equation (1.1). In this way, by means of the univariate conditional probabilities, we can construct an appropriate multivariate distribution for the states of the series.