1 Introduction to Stochastic Processes

1.1 Introduction

Stochastic modelling is an interesting and challenging area of probability and statistics. Our aims in this introductory section of the notes are to explain what a **stochastic process** is and what is meant by the **Markov property**, give examples and discuss some of the objectives that we might have in studying stochastic processes.

1.2 Definitions

We begin with a formal definition, A **stochastic process** is a family of random variables $\{X_\theta\}$, indexed by a parameter θ, where θ belongs to some index set Θ.

In almost all of the examples that we shall look at in this module, Θ will represent time. If Θ is a set of integers, representing specific time points, we have a stochastic process in **discrete time** and we shall replace the general subscript θ by n. So we shall talk about the discrete time process $\{X_n\}$. Sections 2–4 of the module are about processes in discrete time.

If Θ is the real line (or some interval of the real line) we have a stochastic process in **continuous time** and we shall replace the general subscript θ by t and change the notation slightly, writing $X(t)$ rather than X_t. Sections 5–6 of the module are about processes in continuous time.

The reason that we introduce the rather abstract notion of an index set Θ, rather than just working with time, is that we sometimes want to study spatial processes as well as temporal processes. In a spatial process, Θ would be a **vector**, representing location in space rather than time. For example, we might have a process $\{X_{(u,v)}\}$, representing a random variable that varies across two-dimensional space.

Here, $X_{(u,v)}$ represents the value of the process at position (u, v). We can even have processes that evolve in both time and space, so called **spatio-temporal processes**. However, apart from occasional examples, spatial and spatio-temporal processes are beyond the scope of this module.

For processes in time, a less formal definition is that a **stochastic process** is simply a process that develops in time according to probabilistic rules. We shall be particularly concerned with **stationary** processes, in which the probabilistic rules do not change with time.

In general, for a discrete time process, the random variable X_n will depend on earlier values of the process, $X_{n-1}, X_{n-2}, \ldots$. Similarly, in continuous time, $X(t)$ will generally depend on values $X(u)$ for $u < t$.

Therefore, we are often interested in conditional distributions of the form

$$\Pr\left(X_{t_k} | X_{t_{k-1}}, X_{t_{k-2}}, \ldots, X_{t_1}\right)$$

for some set of times $t_k > t_{k-1} > \ldots > t_1$. In general, this conditional distribution will depend upon values of $X_{t_{k-1}}, X_{t_{k-2}}, \ldots, X_{t_1}$. However, we shall focus particularly in this module on processes that satisfy the **Markov** property, which says that

$$\Pr\left(X_{t_k} | X_{t_{k-1}}, X_{t_{k-2}}, \ldots, X_{t_1}\right) = \Pr\left(X_{t_k} | X_{t_{k-1}}\right).$$

The Markov property is named after the Russian probabilist Andrei Andreyevich Markov (1856-1922). An informal mnemonic for remembering the Markov property is this. 'Given the present (X_{k-1}), the future (X_k) is independent of the past $(X_{k-2}, X_{k-3}, \ldots, X_1)$.' The Markov property is sometimes referred to as the 'lack of memory' property.

Stochastic processes that satisfy the Markov property are typically much simpler to analyse than general processes, and most of the processes that we shall study in this module are Markov processes. Of course, in attempting to model any real system it will be important to consider whether the Markov property is likely to hold.

As well as classifying time as discrete or continuous, we can also classify the random variable X as discrete or continuous. We shall see

examples of all four combinations (discrete/continuous time in conjunction with discrete/continuous random variable) in this module.

We end this section with a few more definitions related to stochastic processes:

- A **counting process** is a process $X(t)$ in discrete or continuous time for which the possible values of $X(t)$ are the natural numbers $(0, 1, 2, \ldots)$ with the property that $X(t)$ is a non-decreasing function of t. Often, $X(t)$ can be thought of as counting the number of 'events' of some type that have occurred by time t. The basic example of a counting process is the Poisson process, which we shall study in some detail.
- A **sample path** of a stochastic process is a particular **realisation** of the process, i.e. a particular set of values $X(t)$ for all t (which may be discrete of continuous), generated according to the (stochastic) 'rules' of the process.
- The **increments** of a process are the changes $X(t) - X(s)$ between time points s and t $(s < t)$. Processes in which the increments for non-overlapping time intervals are independent and stationary (i.e. dependent only on the lengths of the time intervals, not the actual times) are of particular importance. Random walks, which we study extensively in Chapter 2, are a good example. General processes of this type are called Lévy processes, and include the Poisson process (Chapter 5) and Brownian motion (Chapter 6).

1.3 Stochastic and deterministic models

Stochastic models can be contrasted with **deterministic** models. A deterministic model is specified by a set of equations that describe **exactly** how the system will evolve over time. In a stochastic model, the evolution is at least partially random and if the process is run several times, it will **not** give identical results. Different runs of a stochastic process are often called **realisations** of the process.

Deterministic models are generally easier to analyse than stochastic models. However, in many cases stochastic models are more realistic, particulary for problems that involve 'small numbers'. For example, suppose we are trying to model the management of a rare species, looking at how different strategies affect the survival of the species. Deterministic models will not be very helpful here, because they will predict that the species either definitely becomes extinct or definitely survives. In a stochastic model, however, there will be a probability of extinction, and we study how this is affected by management practices.

In recent years, the distinction between deterministic and stochastic models has been blurred slightly by the development of chaotic models. A chaotic model is a deterministic model that is extremely sensitive to the values of some of the parameters in the model. Making a very small change to the values of these parameters can make the outcome of the model completely different. Some people have argued that systems that are normally regarded as stochastic processes are better regarded as chaotic deterministic systems, as exemplified by this quote:

A mountain stream, a beating heart, a smallpox epidemic, and a column of rising smoke are all examples of dynamic phenomena that sometimes seem to behave randomly. In actuality, such processes exhibit a special order that scientists and engineers are only just beginning to understand. This special order is 'deterministic chaos', or chaos, for short.

My own, doubtless biased, view is that stochastic models are, in general, much more useful than models based on deterministic chaos. But in any event, chaos theory uses quite different mathematical techniques and is outside the scope of this module.

1.4 Examples of stochastic processes

In this section, we offer an eclectic collection of examples of stochastic processes, to give you some idea of the wide range of application areas.

1.4.1 Exchange rates

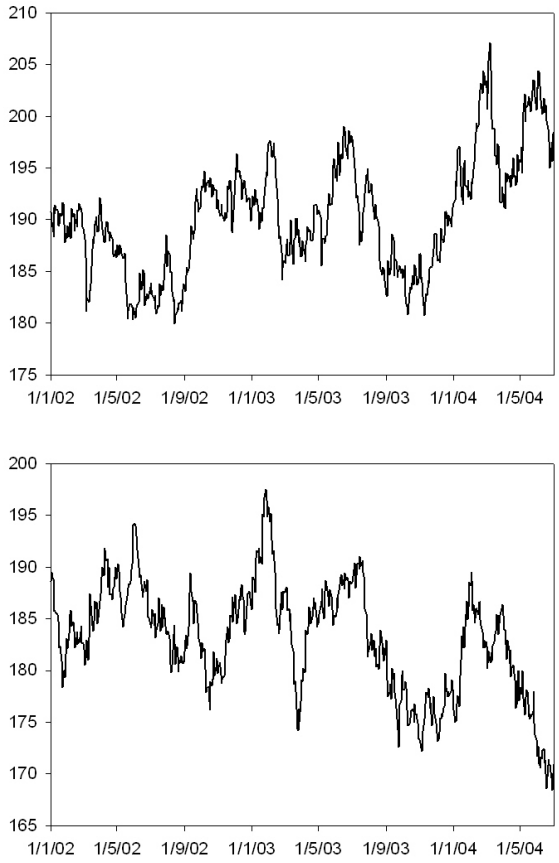

Figure 1: *Exchange rate between British pound and Japanese yen over the period 1/1/02-31/7/04. Upper graph shows true exchange rate, lower graph shows a simulation of a random walk model.*

The upper panel of Figure 1 shows the exchange rate between the British pound and the Japanese yen from 1st January 2002 to 31st July 2004. Over this period, the average exchange rate was 191.1 yen to one pound. The lower panel shows a simulation of a type of stochastic process called a **random walk**. We will be studying random walks in Section 2 of this module. In the random walk model, the daily changes in exchange rate are independent normal random variables with zero mean and standard deviation of 1.206 (matching the

standard deviation in the observed data). Whilst the detailed patterns are of course different, the two series have a similar structure. Note that in the random walk model, upward and downward movements in the exchange rate are equally likely, and there is no scope for making money through currency speculation except by luck.

1.4.2 Photon emission

Photons are minute particles of light. Under some circumstances, a light source will emit photons at random (according to what is called a Poisson process, which we will study in section 5). Individual photons are too faint to be detected by the human eye, but electronic photon detectors *can* detect single photons. However, there is a problem with these machines. Immediately after they have detected a photon, there is a short time period, known as the dead time, during which no new photons can be detected.

The number of photons detected by the machine therefore underestimates the actual number of photons emitted by the light source and we need to correct the observed number of photons to get an estimate of the true number. It's clear that a relatively complicated correction will be needed because when the light source is emitting photons at a low rate, the chances of two photons being emitted close enough together to be affected by dead time is quite small, whereas when the emission rate is high, many photons will be missed. To tackle this problem it is necessary to set up a stochastic model that models both the emission of photons and the dead time effect and then to study the distribution of the observed number of photons and see how this is related to the number actually emitted.

1.4.3 Epidemic models

Figure 2, which is from a World Health Organisation report, shows the progress of the SARS epidemic in 2002-3. The data plotted are the number of new cases reported each week worldwide. The data exclude 2527 probable cases of SARS (mostly from Beijing) in which

the data of onset is unknown. This is an example of a discrete time

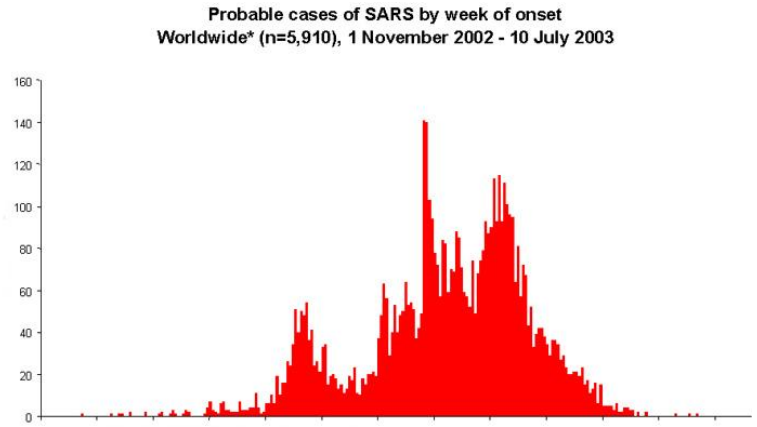

Figure 2: *Daily number of new cases of SARS worldwide during the period 1/11/02–10/7/03. each day*

stochastic process. The variable of interest (number of cases) is also discrete. Many sophisticated mathematical models of epidemics have been developed. These incorporate factors such as the number of contacts that an infected person makes with non-infected people, and the chances that the infection will be passed on during one of these contacts. These models help us to understand the factors that determine whether an epidemic is likely to get a hold and affect a large proportion of the population (like the BSE epidemic in cattle in the UK), or whether it is more likely to affect just a few people and then die out. The models can also be used to predict the effects of interventions - how will it affect the epidemic if we vaccinate 50% of the population?

1.4.4 Earthquakes

Figure 3 shows the occurrences large earthquakes in Alaska during the period 1900-1965. On average, there was roughly one earthquake per year, but in the actual record there are some fairly lengthy periods with no earthquakes, and other periods with several earthquakes in close succession. Is this just what we might expect by chance? Or

does it provide evidence of some clustering of earthquakes in time? If there is clustering, how can we model it? Is clustering related to the magnitudes of the earthquake?

Figure 3: *Times of occurrence of 65 earthquakes of magnitude 7.0 or greater in Alaska, 1900-1965.*

These questions are typical of those that arise when studying **point processes** processes in which we observe the location of events in time or space. For another example, astronomers are interested in the clustering of galaxies, because this can provide clues to the origins of the universe. For earthquakes, some very sophisticated stochastic models of earthquake occurrence have been developed, that incorporate aspects of the underlying geophysics.

1.4.5 Budding Yeast

The yeast species used in brewing and baking, *Saccharomyces cerevisiae* reproduces by budding (Figure 4). The adult cell, which we call the **mother cell**, produces a bud which grows and eventually separates from the mother to produce a new **daughter cell**. The mother cell then produces another bud, after a random time M. The daughter cell, on the other hand, first has to increase in size until it reaches maturity, which takes a random time D. Then it behaves like a mother cell, producing a bud after random time M. Altogether, therefore, the time from when the bud was produced until it produces its own offspring bud is $D + M$.

What can we say about this process? Will the population eventually stabilise in some way, so that, for example, the proportion of daughter cells in the population approaches some fixed value? If so, how does this depend on the distribution of the random variables M and D?

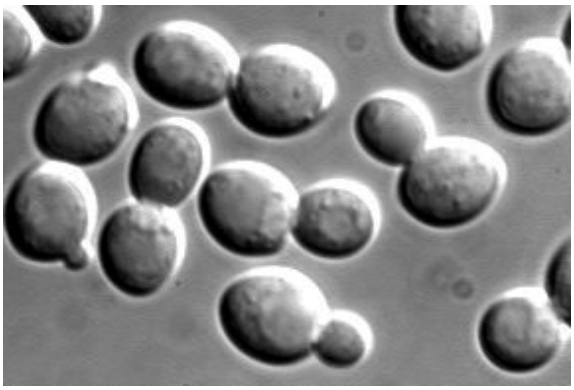

Figure 4: *Electron microscope picture of budding yeast cells.*

Eventually, of course, yeast cells die, after they have produced perhaps 20–40 buds. What effect does this have on the population structure?

1.4.6 Diploid and tetraploid plants

Here we look at one more example in a little more detail. Many plant species are **diploid**, meaning that they carry two copies of each of their genes (like humans). However, some species can also exist as **tetraploids**, with four copies of each gene. The question of how tetraploid plants arise is of great interest to evolutionary biologists.

The following very simple model, based on some genetics that we won't discuss here, mimics some aspects of this process. We have a population of n annual plants, which die at the end of each year, and are replaced by n of their offspring (so the population size remains constant). Initially, all n plants are diploid. In subsequent years, if there are d diploid plants in the current generation, the number of diploid plants in the next generation is a random variable with distribution $\mathrm{Bin}(n, p)$ where

$$p = \frac{u^2 d^2}{u^2 d^2 + (n - ud)^2},$$

and where u is a parameter in the range $(0, 1)$; the larger the value of u, the greater the tendency to produce diploid offsrping plants (and

so the greater the value of p). Notice that this process satisfies the Markov property as the distribution of the number of diploids in the next generation depends only on the number in the current generation.

Figure 5 shows a simulation of this process with a population of $n = 10$ plants and with $u = 0.85$. Eventually, the diploid plants become extinct, but this takes 65 generations. When extinction does occur, it occurs quite rapidly (after 58 generations, the population was still entirely diploid). Note that we wouldn't predict this behaviour if we observed say the first 50 generations only.

This simulation is fairly typical in that there is eventually a fairly rapid extinction of diploids, but the time at which this occurs is very variable. In 10 further simulation runs with the same parameter values, the extinction times were, in ascending order, 12, 17, 18, 22, 33, 70, 72, 93, 102, 315.

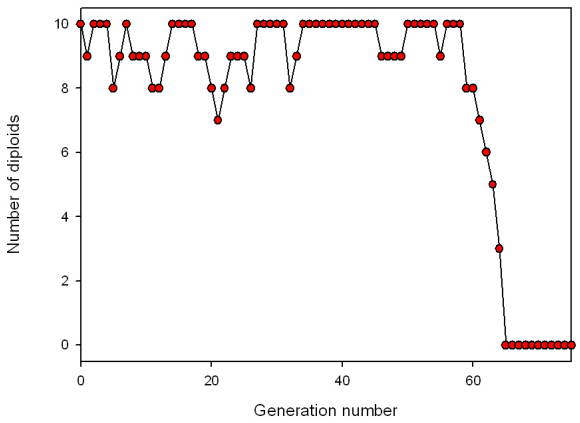

Figure 5: *Simulating the extinction of diploid plants with* $u = 0.85$.

The long period of apparently stable fluctuation before sudden extinction can be seen even more clearly by increasing u to 0.9. Figure 6 shows a simulation of this process. Extinction occurs after 1711 generations, but the population was still entirely diploid after 1705 generations.

In biological terms, the model shows that a population in which the majority of plants are diploid for a long time can nonetheless switch to

being entirely tetraploid over a relatively small number of generations. Although this is a very simplified model, much more complex models, that are more realistic biologically, predict similar behaviour.

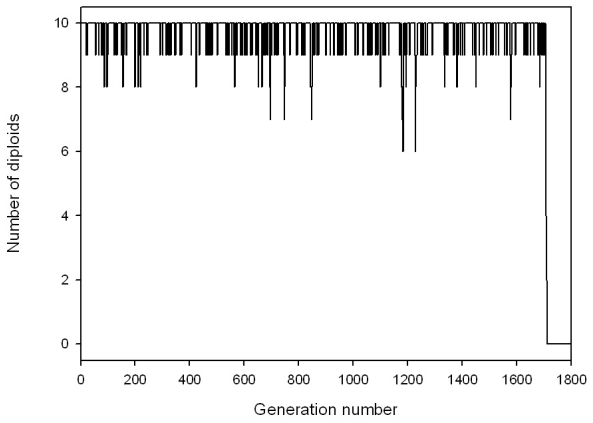

Figure 6: *Simulating the extinction of diploid plants with* $u = 0.9$.

1.5 Modelling

Mathematical models, be they deterministic or stochastic, are intended to mimic real world systems. In particular, they can be used to predict how systems will behave under specified conditions. In scientific work, we may be able to conduct experiments to see if model predictions agree with what actually happens in practice. But in many situations, experimentation is impossible. Even if experimentation is conceivable in principle, it may be impractical for ethical or financial reasons. In these circumstances, the model can only be tested less formally, for example by seeking expert opinion on the predictions of the model.

The examples in the previous section introduce a diverse collection of stochastic models. Others will be presented later in the module. Practical modelling is an art as much as a science. In particular, it's important to try to model at the right level of detail. Too little detail and the model will not be able to make useful predictions; too much and the model may become unwieldly and impractical. Meteorological

modelling provides a good example – short-term forecasting, medium-range forecasting and long-term predictions of climate change involve three quite different modelling approaches.

It's important to remember that 'all models are wrong' in the sense that they are only approximations of reality. What is of interest is whether the model is useful for the purposes for which we want to apply it.

Most real-world models will need to be implemented on a computer. Developing and testing software for models is important but outside the scope of this module. Of course, the fact that a model produces nice computer output does not in itself imply that the model has any validity. Generally a model will need to go through several cycles of improvement before it is finalised. Sometimes, attempting to model a system highlights the fact that there is little information about some aspects of the system and it is necessary to collect new data to try to fill in the gaps.

1.6 Aims of this module

Our aim in this module is to study the basic theory of stochastic processes in discrete and continuous time. We use mathematical techniques to explore the behaviour of these processes. Once these processes are understood, they can be incorporated into real-world models, but this is outside the scope of this particular module.

An alternative way of exploring stochastic processes is to use simulation. Where possible, a mathematical analysis is preferable, because it gives us formulae that explain general behaviour. Simulations have to assume particular values for any unknown parameters in the process. However, simulation is useful for

- Checking analytical results
- Exploring models that are too complex to analyse mathematically

The example of diploid and tetraploid plants shows how we can use simulation to explore a process. In fact, the process described is an example of a Markov chain, and this can be studied analytically, using techniques that we will look at in Section 4. However, if we start to make the model more realistic, by including more biological features, it soon becomes too complex to analyse mathematically and simulation becomes essential.

One thing that we won't be doing in this module, except in rare examples, is looking at how to fit stochastic models to real data. This is complicated by the fact that the data are not usually independent and it can be quite hard, for example, to write down a likelihood. The module MA639: Time Series Modelling has more focus on statistical analysis.