1
Introduction

1.1 Chance and information

Our experience of the world leads us to conclude that many events are unpredictable and sometimes quite unexpected. These may range from the outcome of seemingly simple games such as tossing a coin and trying to guess whether it will be heads or tails to the sudden collapse of governments or the dramatic fall in prices of shares on the stock market. When we try to interpret such events, it is likely that we will take one of two approaches – we will either shrug our shoulders and say it was due to 'chance' or we will argue that we might have have been better able to predict, for example, the government's collapse if only we'd had more 'information' about the machinations of certain ministers. One of the main aims of this book is to demonstrate that these two concepts of 'chance' and 'information' are more closely related than you might think. Indeed, when faced with uncertainty our natural tendency is to search for information that will help us to reduce the uncertainty in our own minds; for example, think of the gambler about to bet on the outcome of a race and combing the sporting papers beforehand for hints about the form of the jockeys and the horses.

Before we proceed further, we should clarify our understanding of the concept of chance. It may be argued that the tossing of fair, unbiased coins is an 'intrinsically random' procedure in that everyone in the world is equally ignorant of whether the result will be heads or tails. On the other hand, our attitude to the fall of governments is a far more subjective business – although you and I might think it extremely unlikely, the prime minister and his or her close advisors will have 'inside information' that guarantees that it's a pretty good bet. Hence, from the point of view of the ordinary citizen of this country, the fall of the government is not the outcome of the play of chance; it only appears that way because of our ignorance of a well-established chain of causation.

Irrespective of the above argument we are going to take the point of view in this book that regards both the tossing of coins and the fall of governments as

falling within the province of 'chance'. To understand the reasoning behind this let us return once again to the fall of the government. Suppose that you are not averse to the occasional flutter and that you are offered the opportunity to bet on the government falling before a certain date. Although events in the corridors of power may already be grinding their way inexorably towards such a conclusion, you are entirely ignorant of these. So, from your point of view, if you decide to bet, then you are taking a chance which may, if you are lucky, lead to you winning some money. This book provides a tool kit for situations such as this one, in which ignorant gamblers are trying to find the best bet in circumstances shrouded by uncertainty.

Formally, this means that we are regarding 'chance' as a relation between individuals and their environment. In fact, the basic starting point of this book will be a person moving through life and encountering various clear-cut 'experiences' such as repeatedly tossing a coin or gambling on the result of a race. So long as the outcome of the experience cannot be predicted in advance by the person experiencing it (even if somebody else can), then chance is at work. This means that we are regarding chance as 'subjective' in the sense that my prediction of whether or not the government will fall may not be the same as that of the prime minister's advisor. Some readers may argue that this means that chance phenomena are unscientific, but this results from a misunderstanding of the scientific endeavour. The aim of science is to obtain a greater understanding of our world. If we find, as we do, that the estimation of chances of events varies from person to person, then our science would be at fault if it failed to reflect this fact.

1.2 Mathematical models of chance phenomena

Let us completely change track and think about a situation that has nothing whatever to do with chance. Suppose that we are planning on building a house and the dimensions of the rectangular base are required to be 50 feet by 25 feet (say). Suppose that we want to know what the lengths of the diagonals are. We would probably go about this by drawing a diagram as shown in Fig. 1.1, and then use Pythagoras' theorem to calculate

$$d = ((50)^2 + (25)^2)^{1/2} = 55.9\,\text{ft}.$$

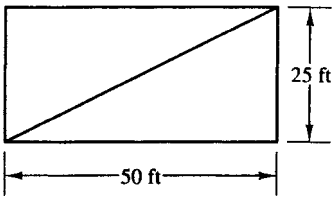

Fig. 1.1.

Let us examine the above chain of reasoning a little more closely. First of all we have taken the walls and floors of a real house that we are proposing to build, which would consist of real bricks and mortar, and have represented these by an abstract drawing consisting of straight lines on paper. Of course, we do this because we know that the precise way in which the walls are built is irrelevant to the calculation we are going to make. We also know that our walls and floorboards do not intersect in exact straight lines but we are happy to use straight lines in our calculation in the knowledge that any errors made are likely to be too tiny to bother us.

Our representation of the floorplan as a rectangle is an example of a *mathematical model* – an abstract representation of part of the world built out of idealised elements.

The next stage of our analysis is the calculation of the diagonal length, and this involves the realisation that there is a *mathematical theory* – in this case, Euclidean geometry – which contains a rich compendium of properties of idealised structures built from straight lines, and which we can use to investigate our particular model. In our case we choose a single result from Euclidean geometry, Pythagoras' theorem, which we can immediately apply to obtain our diagonal length. We should be aware that this number we have calculated is strictly a property of our idealised model and not of a real (or even proposed) house. Nonetheless, the fantastic success rate over the centuries of applying Euclidean geometry in such situations leads us to be highly confident about the correspondence with reality.

The chain of reasoning which we have outlined above is so important that we have highlighted it in Fig. 1.2.

Now let us return to the case of the experience of chance phenomena. We'll consider a very simple example, namely the tossing of a coin which we believe to

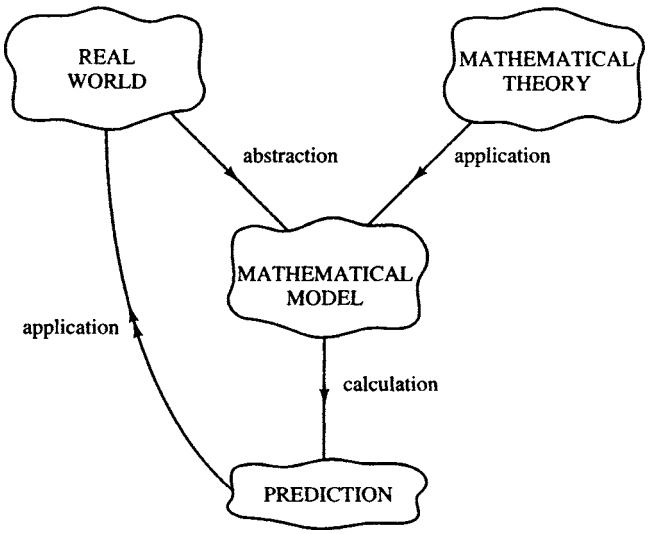

Fig. 1.2.

be fair. Just as it is impossible to find a real straight line in nature, so is it impossible to manufacture a true 'fair' coin – indeed, it is an interesting exercise to speculate how you would test that a coin that is claimed to be perfectly unbiased really is so. I recommend that you think about this question carefully and return to reconsider your answer after you've read Chapter 4. Whether or not you believe in the existence of real fair coins, we are going to consider the behaviour of idealised fair coins as a mathematical model of our real coins. The mathematical theory which then plays the role of Euclidean geometry is called *probability theory* and the development of the basic ideas of this theory is the goal of this book.

One of the main aims of probability theory is, as you might expect, the calculation of probabilities. For example, most of you would agree I'm sure that the probability of a fair coin returning heads is exactly one half. However, although everyone is fairly confident about how to assign probabilities in simple cases like this, there is a great deal of confusion in the literature about what 'probability' means.

We should be aware that probability is a mathematical term which we use to investigate properties of mathematical models of chance phenomema (usually called *probabilistic models*). So 'probability' does not exist out in the real world. Nonetheless, the applications of the subject spread into nearly every corner of modern life. Probability has been successfully applied in every scientific subject (including the social sciences). It has been used to model the mutation of genes, the spread of epidemics (including AIDS) and the changing prices of shares on the stock market. It is the foundation of the science of statistics as well as of statistical mechanics – the physical study of bulk properties of large systems of particles such as gases. We will touch on both these subjects in this book, although our main application will be to use probability to give mathematical meaning to the concept of 'information', which is itself the foundation for the modern theory of communication systems.

The precise definition of probability must wait until Chapter 4, where we will see that it is a kind of generalised notion of 'weight' whereby we weigh events to see how likely they are to occur. The scale runs from 0 to 1, where 1 indicates that an event is certain to happen and 0 that it is impossible. Events with probabilities close to one half are the most uncertain (see Chapter 6).

Just as we develop Pythagoras' theorem in Euclidean geometry and then apply it to mathematical models as we have described above, so we will develop a number of techniques in this book to calculate probabilities, for example if we toss a fair coin five times in a row, by using the binomial distribution (see Chapter 5), we find that the probability of obtaining three heads is $\frac{5}{16}$. If we now want to apply this result to the tossing of real coins, then the situation is somewhat more complicated than in our geometrical example above. The reason for this is of course that 'chance' is a much more complex phenomenon to measure than, say, the length of a wall. In fact, the investigation of the correspondence between chance phenomena in the real

world and the predictions of probabilistic models really belongs to the domain of *statistics* and so is beyond the scope of this book. Here we will be solely concerned with developing methods of calculating probabilities and related concepts, such as averages and entropies, which enable us to analyse probabilistic models as fully as possible. This is the domain of the subject of *probability theory*.

The range of probability theory as we describe it in this book is much wider than that considered by many other authors. Indeed, it is common for textbooks to consider only chance situations which consist of 'scientific' experiments which can be repeated under the same conditions as many times as one wishes. If you want to know the probability of a certain outcome to such an experiment, I'm sure you'll agree that the following procedure will be helpful; that is, you repeat the experiment a large number of times (n, say) and you count the number of incidences of the outcome in question. If this is m, you calculate the *relative frequency* m/n; for example if a coin is tossed 100 times in succession and 60 heads are observed, then the relative frequency of heads is 0.6.

Many mathematicians have attempted to define probability as some kind of limit of relative frequencies, and it can't be denied that such an approach has an appeal. We will discuss this problem in greater detail in Chapters 4 and 8 – for now you may want to think about how such a limit can be calculated in practice. The most rational approach to the problem of relative frequencies is that advocated by the Bayesian school (see Chapter 4). They argue that having made a probabilistic model of a chance experiment, we use all the theoretical means at our disposal to assign *prior probabilities* to all the possible outcomes. We then collect our observations in the form of relative frequencies and use the knowledge gained from these to assign new *posterior probabilities*. So relative frequencies are treated as evidence to be incorporated into probability assignments.

1.3 Mathematical structure and mathematical proof

As probability is a mathematical theory, we need to be clear about how such theories work. A standard way of developing mathematical theories has evolved which goes back to Euclid's geometry. This approach has been developed extensively during the twentieth century and we are going to use it in this book.

First we should note that a *mathematical theory* is a systematic exposition of all the knowledge we possess about a certain area. You may already be familiar with some examples such as set theory or group theory. The essence of a mathematical theory is to begin with some basic definitions, called *axioms*, which describe the main mathematical objects we are interested in, and then use clear logical arguments to deduce the properties of these objects. These new properties are usually announced in statements called *theorems*, and the arguments that we use to convince ourselves of the validity of these theorems are *proofs*. Sometimes it becomes

clear as the theory develops that some new concepts are needed in addition to those given in the axioms, and these are introduced as *definitions*.

In probability theory the basic concept is that of a *probability measure*, for which the axioms are given at the beginning of Chapter 4 (the axioms for the more general concept of measure are given in Chapter 3). One of the most important additional concepts, introduced in Chapter 5, is that of a *random variable*.

There are a number of standard techniques used throughout mathematics for proving theorems. One of the most important is that of proof by mathematical induction. We will use this extensively in the text and if you are not familiar with it you may wish to read Appendix 1. Another useful technique is that of 'proof by contradiction', and we will give a statement and example of how to use this below, just to get you into the swing of things.

Let Q be a proposition that you believe to be true but which you can't prove directly to be true. Let $\sim Q$ be the negation of Q (so that if, for example, Q is the statement 'I am the prime minister', $\sim Q$ is the statement 'I am not the prime minister'). Clearly, either Q or $\sim Q$ (but not both) must hold. The method of the proof is to demonstrate that if $\sim Q$ is valid, then there is a contradiction. Since contradictions are forbidden in mathematics, $\sim Q$ cannot be valid and so Q must be.

In the example given below, Q is the proposition '$\sqrt{2}$ is an irrational number', so that $\sim Q$ is the proposition '$\sqrt{2}$ is a rational number'. We feel free to use the fact that the square root of an even number is always even.

Theorem 1.1 *$\sqrt{2}$ is an irrational number.*

Proof We suppose that $\sqrt{2}$ is rational so we must be able to write it in its lowest terms as

$$\sqrt{2} = \frac{a}{b}.$$

Hence, $a = \sqrt{2}b$ and squaring both sides, $a^2 = 2b^2$, so that a^2 is even and hence a is also even. If a is even, there must be a whole number c (say) such that $a = 2c$ and so $a^2 = 4c^2$.

Substituting for a^2 in the earlier equation $a^2 = 2b^2$ yields $4c^2 = 2b^2$ and so $b^2 = 2c^2$; hence b^2 and also b is even. Thus we can write $b = 2d$ for some whole number d. We now have

$$\sqrt{2} = \frac{a}{b} = \frac{2c}{2d} = \frac{c}{d}.$$

But this contradicts the assumption that $\frac{a}{b}$ was the expression for $\sqrt{2}$ in its lowest terms. □

The symbol □ appearing above is commonly used in mathematics to signify 'end of proof'.

We close this section by listing some additional mathematical nomenclature for statements:

Lemma – this is usually a minor technical result which may be a stepping stone towards a theorem.

Proposition – in between a lemma and a theorem. Sometimes it indicates a theorem from a different branch of mathematics, which is needed so that it can be applied within the current theory.

Corollary – a result that follows almost immediately from the theorem with very little additional argument.

1.4 Plan of this book

This is an introductory account of some of the basic ideas of probability theory and information theory. The only prerequisites for reading it are a reasonable ability at algebraic manipulation and having mastered a standard introductory course in the calculus of a single variable, although calculus is not used too often in the first seven chapters. The main exception to this is the extensive use of partial differentiation and, specifically, Lagrange multipliers in Section 6.4, but if you are not familiar with these, you should first read Appendix 2 at the end of the book. You should also brush up your knowledge of the properties of logarithms before starting Chapter 6. I have tried to avoid any use of rigorous mathematical analysis, but some sort of idea of the notion of a limit (even if only an intuitive one) will be helpful. In particular, if you find the discussion of integration in Section 8.3 too difficult, you can leave it and all subsequent references to it without any great loss. For Chapter 9 you will need to know the rudiments of double integration. Chapter 10 requires some knowledge of matrix algebra and all of the material that you need from this area is reviewed in Appendix 5. Two sections of the book, Sections 6.6 and 7.5, are somewhat more difficult than the rest of the book and you may want to skip these at the first reading.

At the end of each chapter you will find a set of exercises to work through. These days many textbooks carry the health warning that 'the exercises are an integral part of the text' and this book is no exception – indeed, many results are used freely in the text that you are invited to prove for yourself in the exercises. Solutions to numerical exercises and some of the more important theoretical ones can be found at the end of the book. Exercises marked with a $(*)$ are harder than average; you may wish to skip these (and any other starred Section) at the first reading. You will also find at the end of each chapter some guidance towards further reading if you want to explore some of the themes in greater detail.

Now a brief tour through the book. Chapter 3 describes a number of counting tricks that are very useful in solving probabilistic problems. In Chapter 3, we give a brief account of set theory and Boolean algebra, which are the modern context of probability theory. In particular, we learn how to 'measure' the 'weight' of a set. In Chapter 4, we find that this measuring technique is precisely the mathematical tool

we need to describe the probability of an event. We also learn about conditioning and independence and survey some of the competing interpretations of probability. Discrete random variables are introduced in Chapter 5, along with their properties of expectation and variance. Examples include Bernoulli, binomial and Poisson random variables.

The concepts of information and entropy are studied in Chapter 6. Entropy is one of the most deep and fascinating concepts in mathematics. It was first introduced as a measure of disorder in physical systems, but for us it will be most important in a dual role as representing average information and degree of uncertainty. We will present the maximum entropy principle, which employs entropy as a tool in selecting (prior) probability distributions. Chapter 7 applies information theoretic concepts to the study of simple models of communication. We investigate the effects of coding on the transmission of information and prove (in a simple case) Shannon's fundamental theorem on the (theoretical) conditions for optimal transmission.

In the next two chapters we generalise to random variables with continuous ranges. In particular, in Chapter 8 we establish the weak law of large numbers, examine the normal distribution and go on to prove the central limit theorem (perhaps the most important result in the book). We also examine the continuous analogue of entropy. Random vectors and their (multivariate) distributions are studied in Chapter 9 and we use these to investigate conditional density functions. We are then able to analyse a simple model of the communication of continuous signals. So far all of the theoretical development and modelling has been 'static' in that there has been no attempt to describe the passing of time. Chapter 10 addresses this problem by introducing (discrete-time) Markov chains, which form an important class of random processes. We study these from both the probabilistic and information theoretic viewpoints and one of the highlights is the derivation of a very attractive and concise formula for the entropy rate of a stationary Markov chain. Some readers may feel that they already know about probability and want to dive straight into the information. They should turn straight to Chapters 6 and 7 and then study Sections 8.7, 9.6, 9.7 and 10.6.

The concept of probability, which we develop in this book, is not the most general one. Firstly, we use Boolean algebras rather than σ-algebras to describe events. This is a technical restriction which is designed to make it easier for you to learn the subject, and you shouldn't worry too much about it; more details for those who want them are given at the end of Chapter 4. Secondly and more interestingly, when we descend into the microscopic world of atoms, molecules and more exotic particles, where nature reveals itself sometimes as 'particles' and other times as 'waves', we find that our observations are even more widely ruled by chance than those in the everyday world. However, just as the classical mechanics of Newton is no longer appropriate to the description of the physics in this landscape, and

we have instead to use the strange laws of quantum mechanics, so the 'classical' probability we develop in this book is no longer adequate here and in its place we must use 'quantum probability'. Although this is a rapidly growing and fascinating subject, it requires knowledge of a great deal of modern mathematics, which is far beyond the scope of this book and so must be postponed by the interested reader for later study.

2
Combinatorics

2.1 Counting

This chapter will be devoted to problems involving counting. Of course, everybody knows how to count, but sometimes this can be quite a tricky business. Consider, for example, the following questions:

(i) In how many different ways can seven identical objects be arranged in a row?
(ii) In how many different ways can a group of three ball bearings be selected from a bag containing eight?

Problems of this type are called *combinatorial*. If you try to solve them directly by counting all the possible alternatives, you will find this to be a laborious and time-consuming procedure. Fortunately, a number of clever tricks are available which save you from having to do this. The branch of mathematics which develops these is called *combinatorics* and the purpose of the present chapter is to give a brief introduction to this topic.

A fundamental concept both in this chapter and the subsequent ones on probability theory proper will be that of an 'experience' which can result in several possible 'outcomes'. Examples of such experiences are:

(a) throwing a die where the possible outcomes are the six faces which can appear,
(b) queueing at a bus-stop where the outcomes consist of the nine different buses, serving different routes, which stop there.

If A and B are two separate experiences, we write $A \circ B$ to denote the combined experience of A followed by B. So if we combine the two experiences in the examples above, we will find that $A \circ B$ is the experience of first throwing a die and then waiting for a bus. A natural question to ask is how many outcomes there are in $A \circ B$. This is answered by the following result.