colleagues have had comments and
esag, Michael Phelan, and Elizabeth
as well as a host of students over the
helpful comments: Simon Tavaré,
aurence Baxter. John Kimmel was a
ıe project, while Achi Dosanjh and
rough the final stages of manuscript
erously provided data, figures or other
João Batista, David Brillinger, Anna
phen Kaluzny, Niels Keiding, Brian
Roland Madden, Thomas Murray,
ın Rice, Brian Ripley, Nuala Sheehan,

from Elan Computer Group, Inc. The
Splus (StatSci division of Mathsoft,
y from the psfig utility written by
written by Frank Harrell. The spatial
cs library produced by B. S. Rowling-
ersity. Software for some specialized
nd John Rice.

om the National Science Foundation
Health (HL 31823), the Environmen-
Power Research Institute. This text
policy of any of these organizations.
a visiting scholar appointment at the
e University of Washington, and I am
a most difficult time.

rough my long preoccupation with this
much more attention (and will, hope-
y supportive of the effort.

man and David Brillinger as advisers
I learned a way of thinking about sci-
nd I learned, in Mr. Neyman's words,
esting."

Peter Guttorp
University of Washington
Seattle 1995

CHAPTER 1

Introduction

Random behavior of simple or complex phenomena can sometimes be explained in physical terms, with an additional touch of probability theory. We exemplify this with a description of coin tossing. We then define a stochastic process, give some examples, and an overview of the book.

1.1. Randomness

The world is full of unpredictable events. Science strives to understand natural phenomena, in the sense of reducing this unpredictability. There are many ways of doing this. Models, which are abstractions of particular aspects of phenomena, constitute one such way. Experimentation and observation are needed to verify and improve models. These models can be of a variety of types: conceptual, mathematical, or stochastic, to mention a few.

In this book we investigate some simple stochastic models. We shall see how there is an interplay between the model and the science underlying the problem. Sometimes the science enters only conceptually, while in other cases it dictates the precise structure of the model. Common to all models we shall consider is a random element, described very precisely in the language of probability theory.

We can qualitatively distinguish different sources of random behavior.

- *Uncertainty about initial conditions.* In many situations it is very difficult to determine exactly the initial conditions. In some situations one can only determine relative frequencies of different initial conditions. The consequence is that the system exhibits random behavior, in accordance with the random initial conditions.

- *Sensitivity to initial conditions.* Many systems exhibit large changes in output corresponding to very small changes in initial conditions. Such systems are said to display **chaotic** behavior. It is often quite difficult to estimate parameters and study the fit of a deterministic description of a chaotic system, especially when initial conditions are not exactly known.

• *Incomplete description.* Sometimes the theoretical basis for a deterministic description of a system corresponds to only some of the factors that are important in determining outcomes. The lack of complete description renders the behavior unpredictable. This is quite common in, e.g., economic modeling.

• *Fundamental description.* When a single photon hits a glass surface, it may go through it or reflect off it. There is no way to predict what it will do, but the quantum electrodynamic theory can make precise predictions as to the behavior of a large number of single photons, when viewed as an aggregate. Modern quantum theory could not exist without probabilistic descriptions.

To illustrate the first two classes of randomness, let us discuss the simple phenomenon of coin tossing. A circular coin of radius a is tossed upwards, spins around its own axis several times, and falls down. For simplicity we will ignore air resistance, and assume that the coin falls on a sandy surface, so that it does not bounce once it hits the ground. The laws of motion for coin tossing are then elementary (Keller, 1986). Let $y(t)$ be the height of the center of gravity of the coin above the surface at time t, and $\theta(t)$ the angle between the normal to the surface and the normal to the heads side of the coin. Figure 1.1 shows the situation.

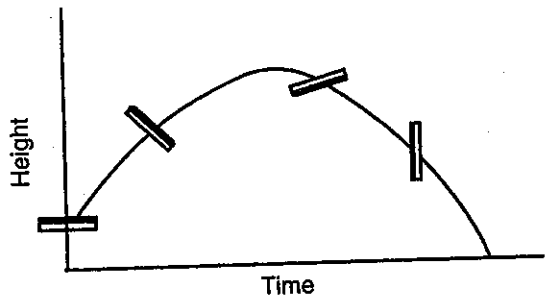


Figure 1.1. The motion of a coin flip. Adapted from Keller (1986); reproduced with permission from the Mathematical Association of America.

Assuming only gravitational force vertically and no acceleration rotationally, the equations of motion are

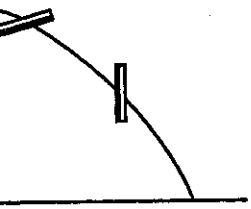
$$y''(t) = -g \quad \theta''(t)=0, \tag{1.1}$$

where g is the gravitational acceleration. The coin is assumed to start at a height a above the surface (any other initial position can be assumed without problems) with the heads side facing upwards. It is tossed upwards with vertical velocity u and angular velocity ω. This yields the initial conditions

es the theoretical basis for a deter-
responds to only some of the factors
outcomes. The lack of complete
redictable. This is quite common in,

single photon hits a glass surface, it
here is no way to predict what it will
theory can make precise predictions
of single photons, when viewed as an
could not exist without probabilistic

randomness, let us discuss the simple
in of radius a is tossed upwards, spins
s down. For simplicity we will ignore
lls on a sandy surface, so that it does
ws of motion for coin tossing are then
height of the center of gravity of the
the angle between the normal to the
f the coin. Figure 1.1 shows the situa-



ne

n flip. Adapted from Keller (1986);
ematical Association of America.

ally and no acceleration rotationally,

(1.1)

n. The coin is assumed to start at a
itial position can be assumed without
ards. It is tossed upwards with vertical
ields the initial conditions

$$y(0)=a \quad y'(0)=u \quad \theta(0)=0 \quad \theta'(0)=\omega. \tag{1.2}$$

The solution is

$$y(t) = a + ut - gt^2/2 \qquad \theta(t) = \omega t. \tag{1.3}$$

The center of the coin describes a parabola as a function of time, with a maximum value of $a+u^2/(2g)$ at time u/g. The coin lands whenever either end touches the ground, i.e., at the first time t_0 such that

$$y(t_0) = a\,|\sin\theta(t_0)|. \tag{1.4}$$

It lands with heads up if it has rotated any number of full rotations (to within 90°), i.e., if for some n

$$(2n-\tfrac{1}{2})\pi < \theta(t_0) < (2n+\tfrac{1}{2})\pi. \tag{1.5}$$

At the extreme points of these intervals $\omega t_0=(2n\pm\frac{1}{2})\pi$, so $y(t_0)=a$ or $t_0=2u/g$. Hence, using (1.3), $\omega=(2n\pm\frac{1}{2})\pi g/(2u)$. The initial values that ensure heads are contained in successive hyperbolic bands:

$$H = \{(u,\omega)\colon (2n-\tfrac{1}{2})\frac{\pi g}{2u} \le \omega \le (2n+\tfrac{1}{2})\frac{\pi g}{2u}\}. \tag{1.6}$$

Figure 1.2 shows the successive bands as a function of u/g. The band nearest the origin corresponds to heads, the next one to tails, etc. We see that as the velocities increase, the amount of change needed to move from tails to heads decreases. Typical values of the parameters were determined in an experiment by Persi Diaconis (Keller, 1986). They are

u = 2.4 m/s (determined from observing the maximum height of typical coin tosses);
ω = 38 rev/s = 238.6 radians/s (determined by wrapping dental floss around the coin and counting the number of unwrappings);
n = 19 rev/toss (computed from $n=\omega t_0$).

On Figure 1.2 this corresponds to u/g of 0.25, and ω of 238.6, far above the top of the picture. It is clear that a very small change in u/g can change the outcome of the toss. Thus the coin toss is predictable precisely when the initial conditions are very well determined. For example, if u/g and ω both are very small, the coin will essentially just fall flat down, and hence will come up heads.

Suppose now that the initial conditions are not under perfect control, but that they can be described by a probability density $f(u,w)$. Then

$$\mathbf{P}(H) = \int_H f(u,\omega)\,du\,d\omega, \tag{1.7}$$

where H is the set depicted in Figure 1.2. We can select f so that $\mathbf{P}(H)$ takes on any value in [0,1]. So how come coins tend to come up heads and tails about equally often? Here is an explanation:

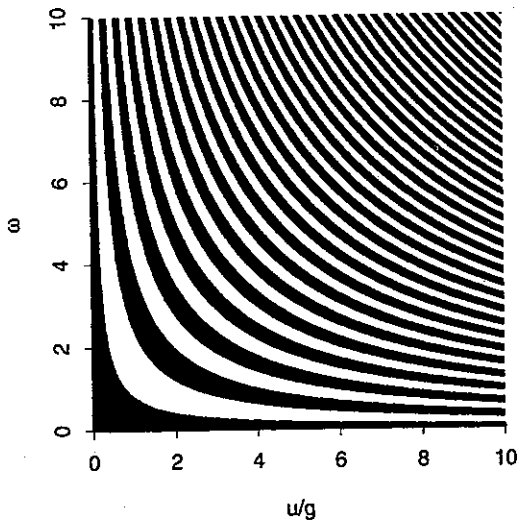


Figure 1.2. The outcome of a coin toss as a function of vertical (u/g) and angular (ω) velocities. Every other band is heads, starting with the lower left-hand corner. Reproduced with permission from Keller (1986), published by the Mathematical Association of America.

Theorem 1.1 For all continuous f such that $f(u,\omega)>0$ for all $u,\omega>0$, and for any $0<\beta\leq\pi/2$

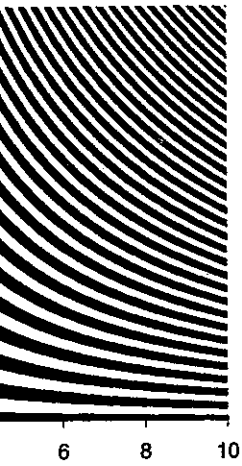
$$\lim_{U\to\infty}\iint_H f(u-U\cos\beta,\omega-\frac{U}{a}\sin\beta)d\omega du = \tfrac{1}{2}. \tag{1.8}$$

Proof First change variables to $\omega'=\omega-U\sin\beta/a$, so that $d\omega'=d\omega$ and

$$\mathbf{P}(H) = \int_{U\cos\beta}^{\infty}\sum_{n=0}^{\infty}\int_{(2n-\frac{1}{2})\pi g/2u-U\sin\beta/a}^{(2n+\frac{1}{2})\pi g/2u-U\sin\beta/a} f(u-U\cos\beta,\omega')d\omega' du. \tag{1.9}$$

For $\beta<\pi/2$ ($\beta=\pi/2$ is argued similarly), $U\to\infty$ implies that $u\to\infty$. The range of each of the inner integrals over ω' is of length $O(1/u)$[1], so we can approximate each of them by the length of the interval times the integrand at the midpoint, with an error of only $o(u^{-1})$. Thus

[1] We say that $g(s)=O(h(s))$ as $s\to s_0$ if $|g(s)/h(s)|$ stays bounded as $s\to s_0$. Here s_0 can be finite or infinite, and the limit can be through all reals or through integers. Also, we write $g(s)=o(h(s))$ if $g(s)/h(s)\to 0$. This notation is due to Landau (1930).



u/g

ss as a function of vertical (u/g) and
is heads, starting with the lower left-
from Keller (1986), published by the

ch that $f(u,\omega)>0$ for all $u,\omega>0$, and

$$)d\omega du = \tfrac{1}{2}. \tag{1.8}$$

$-U\sin\beta/a$, so that $d\omega'=d\omega$ and

$$f(u-U\cos\beta,\omega')d\omega' du. \tag{1.9}$$

$\to\infty$ implies that $u\to\infty$. The range of
ngth $O(1/u)$[1], so we can approximate
l times the integrand at the midpoint,

| stays bounded as $s\to s_0$. Here s_0 can be finite
hrough integers. Also, we write $g(s)=o(h(s))$ if

$$\mathbf{P}(H) = \int_{U\cos\beta}^{\infty} \sum_{n=0}^{\infty} f(u-U\cos\beta, \frac{2n\pi g}{2u}) \frac{\pi g}{2u}(1+o(1))du. \tag{1.10}$$

For large U the sum is a Riemann sum for $\frac{1}{2}\int f(u-U\cos\beta,w)dw$, so with $v=u-U\cos\beta$ we get

$$\mathbf{P}(H) = \tfrac{1}{2}\int_0^{\infty}\int_0^{\infty} f(v,w)dwdv(1+o(1)) = \tfrac{1}{2}+o(1), \tag{1.11}$$

proving the result. □

The interpretation of this theorem is that as soon as we make the initial velocity (vertical, angular, or both) large enough, the probability of heads is ½. One would be tempted to argue this result from symmetry, but the initial conditions are not symmetric: we always start with heads upwards.

More precise results of this type are given in section 3.2.2 of Engel (1992). For example, if the vertical velocity $u\sim$U(2.1,2.7) m/s and the angular velocity $\omega\sim$U(36,40) rev/s, with the two velocities independent, then $|\mathbf{P}(H)-\frac{1}{2}|\le 0.056$. This is, in essence, a worst-case scenario for the approximation. More realistically, assume that the conditional distribution of u given ω is a mixture of normal distributions, with standard deviations at least 0.15 m/s, while ω is at least 36 rev/s. Then Engel's bound is $|\mathbf{P}(H)-\frac{1}{2}|\le 1.5\times 10^{-11}$.

In practice, it is difficult to study processes of the coin-tossing type. One of the most extensive experiments involved 315,672 throws of dice. These were rolled, twelve at a time, down a fixed slope of cardboard. The roller, W. F. R. Weldon (see Pearson, 1900), found 106,602 instances of the outcome 5 or 6. If the dice were true and the initial conditions "sufficiently random", as discussed above, the probability of 5 or 6 should be 1/3. Thus, Weldon found an excess of 0.004366, which is statistically highly significant. The explanation is that the faces with 5 or 6 spots are lighter, because of the larger number of pits for the marking material, thus displacing the center of gravity towards the opposing sides 2 or 1 (this explanation is due to Jeffreys, 1939). In order to avoid this type of a problem, dice would have to have painted faces rather than the pits that are standard. This will, of course, be of little importance to the occasional board game player, but can play an important role in the finances of a major casino.

1.2. Stochastic processes

Classical statistical theory is often concerned with the case of iid random variables $X_0,\ldots,X_n$, i.e.,

$$\mathbf{P}(X_0\in A_0,\ldots,X_n\in A_n) = \prod_{i=0}^{n}\mathbf{P}(X\in A_i), \tag{1.12}$$

where X is a generic random variable with the same distribution as the X_i. It is straightforward to relax the assumption of identical distributions. But in many

cases the independence assumption is violated. For example, if it is raining today, it is more likely to be raining tomorrow than if it is sunny today. There are many different ways of specifying dependent random variables, and we shall see some of them in this book. A **stochastic process** is a collection of random variables $(X_\alpha;\alpha\in T)$ where T is some index set. A formula similar to (1.12) that holds regardless of the dependence between the variables is the following:

$$\mathbf{P}(X_0\in A_0,\ldots,X_n\in A_n)$$
$$=\mathbf{P}\{X_0\in A\}\prod_{i=1}^{n}\mathbf{P}(X_i\in A_i\,|\,X_0\in A_0,\ldots,X_{i-1}\in A_{i-1}). \tag{1.13}$$

The theory of stochastic processes provides various specifications of the conditional probabilities on the right-hand side of (1.13). We will often relate such specifications to natural processes, the outcome of which we may be uncertain about. We will be concerned with random variables taking values in a **state space**[1] S, and governed by a probability measure that we call $\mathbf{P}$. Not all specifications of the type (1.13) are necessarily valid. Various kinds of regularity conditions are needed. We give two different examples of such regularity next.

The positivity condition. Consider a discrete multivariate random variable $\mathbf{X}=(X_1,\ldots,X_m)$ with probability distribution $q(\mathbf{x})=\mathbf{P}(\mathbf{X}=\mathbf{x})$, $\mathbf{x}\in S\equiv\{\mathbf{x}:q(\mathbf{x})>0\}$. This choice of S is called the **minimal state space**. Now consider the minimal state spaces for each of the components X_i, i.e., $S_i=\{x:\mathbf{P}(X_i=x)>0\}$. We say that q satisfies the **positivity condition** if

$$S=S_1\times S_2\times\cdots\times S_m, \tag{1.14}$$

so that if $x_i\in S_i$, $i=1,\ldots,m$, we have $\mathbf{x}=(x_1,\ldots,x_m)\in S$.

Example (The positivity condition) Let $m=2$ and consider binary X_i. If the X_i are independently distributed as Bin(1,½) we have

$$q(0,0)=q(1,1)=q(0,1)=q(1,0)=1/4 \tag{1.15}$$

which does satisfy the positivity condition. On the other hand, if

$$q(0,0)=q(1,1)=½ \tag{1.16}$$

we have $S_i=\{0,1\}$, so

$$S_1\times S_2=((0,0),(0,1),(1,0),(1,1)) \tag{1.17}$$

while

$$S=((0,0),(1,1)), \tag{1.18}$$

[1]The state space is often called the sample space in introductory probability. We use the term state space which is more common in stochastic process theory.

plated. For example, if it is raining
rrow than if it is sunny today. There
ndent random variables, and we shall
tic process is a collection of random
set. A formula similar to (1.12) that
n the variables is the following:

$X_0 \in A_0, \ldots, X_{i-1} \in A_{i-1}).$ (1.13)

es various specifications of the condi-
of (1.13). We will often relate such
tcome of which we may be uncertain
m variables taking values in a **state**
y measure that we call **P**. Not all
sarily valid. Various kinds of regular-
different examples of such regularity

liscrete multivariate random variable
distribution $q(\mathbf{x}) = \mathbf{P}(\mathbf{X}=\mathbf{x})$,
alled the **minimal state space**. Now
each of the components X_i, i.e.,
es the **positivity condition** if

(1.14)

$x_1, \ldots, x_m) \in S$.

) Let $m = 2$ and consider binary X_i.
Bin(1,½) we have

(1,0) = 1/4 (1.15)

. On the other hand, if

(1.16)

1,1)) (1.17)

(1.18)

introductory probability. We use the term state
theory.

violating the condition. □

The conditional distribution of X_i, given the values of all the other variables which we denote $\mathbf{X}_{-i}$, is

$$q_i(x_i \mid \mathbf{x}_{-i}) = \mathbf{P}(X_i=x_i \mid \mathbf{X}_{-i}=\mathbf{x}_{-i}) = \frac{\mathbf{P}(\mathbf{X}=\mathbf{x})}{\mathbf{P}(\mathbf{X}_{-i}=\mathbf{x}_{-i})} = \frac{q(\mathbf{x})}{\sum_{x_i \in S_i} q(\mathbf{x})}. \quad (1.19)$$

Clearly, if q satisfies the positivity condition these q_i are well defined and positive for any $\mathbf{x} \in S$. Note that the q_i are univariate probability distributions. It turns out that under the positivity condition knowledge of the q_i suffices to determine q. We shall see uses of this later on. Let $\mathbf{x}_s^t$ be shorthand for $x_s, \ldots, x_t$, interpreted as an empty set if $s>t$. The following expansion is due to Brook (1964).

Proposition 1.1 Suppose that q satisfies the positivity condition. Then

$$\frac{q(\mathbf{x})}{q(\mathbf{y})} = \prod_{i=1}^{m} \frac{q_i(x_i \mid \mathbf{x}_1^{i-1}, \mathbf{y}_{i+1}^m)}{q_i(y_i \mid \mathbf{x}_1^{i-1}, \mathbf{y}_{i+1}^m)}. \quad (1.20)$$

Proof

$$\begin{aligned} q(\mathbf{x}) &= q_m(x_m \mid \mathbf{x}_1^{m-1}) \mathbf{P}(\mathbf{X}_1^{m-1}=\mathbf{x}_1^{m-1}) \\ &= q_m(x_m \mid \mathbf{x}_1^{m-1}) \mathbf{P}(\mathbf{X}_1^{m-1}=\mathbf{x}_1^{m-1}) \frac{q_m(y_m \mid \mathbf{x}_1^{m-1})}{q_m(y_m \mid \mathbf{x}_1^{m-1})} \\ &= \frac{q_m(x_m \mid \mathbf{x}_1^{m-1})}{q_m(y_m \mid \mathbf{x}_1^{m-1})} q(\mathbf{x}_1^m, y_m) \quad (1.21) \\ &= \frac{q_m(x_m \mid \mathbf{x}_1^{m-1})}{q_m(y_m \mid \mathbf{x}_1^{m-1})} q_{m-1}(x_{m-1} \mid \mathbf{x}_1^{m-2}, y_m) \mathbf{P}(\mathbf{X}_1^{m-2}=\mathbf{x}_1^{m-2}, X_m=y_m) \\ &= \frac{q_m(x_m \mid \mathbf{x}_1^{m-1})}{q_m(y_m \mid \mathbf{x}_1^{m-1})} \frac{q_{m-1}(x_{m-1} \mid \mathbf{x}_1^{m-2}, y_m)}{q_{m-1}(y_{m-1} \mid \mathbf{x}_1^{m-2}, y_m)} q(\mathbf{x}_1^{m-2}, \mathbf{y}_{m-1}^m). \end{aligned}$$

Continue this process for each i, "replacing" x_i with y_i by multiplying and dividing by $q_i(y_i \mid \mathbf{x}_1^{i-1}, \mathbf{y}_{i+1}^m)$ and regrouping terms. Positivity assures that if $q(\mathbf{x})>0$ and $q(\mathbf{y})>0$ then $q(\mathbf{x}_1^{i-1}, \mathbf{y}_i^m)>0$ for $i=1, \ldots, m$ (Exercise 1). □

The consequence of this expansion is that q is uniquely determined by the conditional distributions, since $\sum_{\mathbf{x} \in S} q(\mathbf{x})=1$.

The Kolmogorov consistency condition. Consider a stochastic process $(X_i; i=0,1,...)$, having a distribution such that

$$\mathbf{P}(X_{i_1} \in A_1, \ldots, X_{i_n} \in A_n) = \mathbf{P}(X_{i_1} \in A_1, \ldots, X_{i_n} \in A_n, X_{i_{n+1}} \in S) \qquad (1.22)$$

for all $i_1, \ldots, i_{n+1} \in \{0,1,...rb$, $n \in \mathbf{Z}_+$, and $A_1, \ldots, A_n$ events (measurable subsets of S). This condition guarantees the existence of a probability measure corresponding to this stochastic process, and is due to Kolmogorov[1] (1933).

Remark It is nontrivial that any stochastic processes exist, in the sense of allowing the description of statments about it in terms of probabilities. In fact, one must impose some structure on S. This is the subject of extension theorems in measure-theoretic probability theory. □

Example (An iid process) The standard iid model so common in probability and statistics satisfies the Kolmogorov condition. In the case of random variables with density $f(x)$, taking values on the real line, so $S=\mathbf{R}$, we have

$$\begin{aligned}
&\mathbf{P}(X_{\alpha_1} \in A_1, \ldots, X_{\alpha_n} \in A_n, X_{\alpha_{n+1}} \in \mathbf{R}) \\
&= \mathbf{P}(X_{\alpha_1} \in A_1) \cdots \mathbf{P}(X_{\alpha_n} \in A_n) \mathbf{P}(X_{\alpha_{n+1}} \in \mathbf{R}) \\
&= \mathbf{P}(X_{\alpha_1} \in A_1) \cdots \mathbf{P}(X_{\alpha_n} \in A_n) \int_{-\infty}^{\infty} f(x)dx \qquad (1.23) \\
&= \mathbf{P}(X_{\alpha_1} \in A_1) \cdots \mathbf{P}(X_{\alpha_n} \in A_n) = \mathbf{P}(X_{\alpha_1} \in A_1, \ldots, X_{\alpha_n} \in A_n).
\end{aligned}$$

□

Depending on the structure of the state space S and the index set T one has different classifications of stochastic processes. T is often called "time" in a generic sense.

Example (Some interpretations of time) 1. X_t is the number of earthquakes of magnitude above 5 near Mount St. Helens in the time period $(0,t]$ where 0 is the beginning of recording. This is called a **counting process.** We have $T=\mathbf{R}_+$ and $S=\mathbf{N}$. Time is **continuous** and the state space is **discrete.**

2. $X_k=(B_k,D_k)$ are the number of births and deaths on day k in a population of insects. Here $T=\mathbb{N}$ and $S=\mathbb{N}^2$. Time and state space are both **discrete.**

[1] Kolmogorov, Andrei Nikolaievich (1903–1987). Russian probabilist of the Moscow School, student of Luzin. Developed the axiomatization of probability theory. Made immense contributions to the theory of stochastic processes, turbulence, and complexity theory.

ion. Consider a stochastic process
ıat

$$X_{i_1}\in A_1,\ldots,X_{i_n}\in A_n,X_{i_{n+1}}\in S) \quad (1.22)$$

$A_1,\ldots,A_n$ events (measurable sub-
e existence of a probability measure
nd is due to Kolmogorov[1] (1933).

hastic processes exist, in the sense of
ut it in terms of probabilities. In fact,
is is the subject of extension theorems
□

ndard iid model so common in proba-
orov condition. In the case of random
on the real line, so $S=\mathbf{R}$, we have

$\in\mathbf{R}$)

$P(X_{\alpha_{n+1}}\in\mathbf{R})$

$$\int_{-\infty}^{\infty} f(x)dx \quad (1.23)$$

$= \mathbf{P}(X_{\alpha_1}\in A_1,\ldots,X_{\alpha_n}\in A_n).$

□

state space S and the index set T one
processes. T is often called "time" in a

f time) 1. X_t is the number of earth-
nt St. Helens in the time period $(0,t]$
This is called a **counting process**. We
is and the state space is **discrete**.

nd deaths on day k in a population of
state space are both **discrete**.

). Russian probabilist of the Moscow School,
ion of probability theory. Made immense
turbulence, and complexity theory.

3. $X_{y,t}$ is the amount of SO_4^{2-} in precipitation at location y at time t. Here $T=\mathbf{R}^2\times\mathbf{R}_+$ (or some appropriate subset thereof), $\alpha=(y,t)$, and $S=[0,\infty)$. This is called a **random field** (because T is more than one-dimensional). The state space is **continuous**. Here clock time is only one component of "time".

4. X_t is the thickness of an optical fiber at a distance t from the origin. Here both state space and time are **continuous** with $T=S=\mathbf{R}_+$. "Time" really is distance.
□

Much of the history of the subject of stochastic processes is rooted in particular physical, biological, social, or medical phenomena. The first occurrence of what is now called a Markov chain may have been as an alternative to the simple iid model in explaining rainfall patterns in Brussels (Quetelet, 1852). The simple branching process was invented by Bienaymé (1845) to compute the probability of extinction of a family surname among nobility. Rutherford and Geiger (1910) enrolled the help of the mathematician H. Bateman to describe the disintegration of radioactive substances using what we now call a Poisson process. Einstein (1905) presented a stochastic process that described well the Brownian motion of gold particles in solution, and Bachelier (1900) had used the same process to describe bond prices. The birth-and-death process was introduced by McKendrick (1924; in a special case in 1914) to describe epidemics, and Gibbs (1902) used nearest-neighbor interaction models to describe the behavior of large systems of molecules. The literature lacks a systematic account of the history of stochastic processes, but it should be clear from this short list of examples that an important aspect of such a history is the scientific problems that generated these different stochastic processes. The historical development of the probability theory associated with these processes rapidly moves beyond the scope of this book.

1.3. Purposes of stochastic models

The use of statistical methods to draw scientific conclusions will be illustrated repeatedly in this text. As a first, and quite simple, example, consider the question of mutation in bacteria. It was well known before 1943 that a pure culture of bacteria could give rise to a small number of cells exhibiting different, and inheritable, behavior. For example, when a plate of *E. coli* becomes infected by the bacteriophage T1, most of the cells are destroyed. A few may, however, survive. These, as well as all their offspring, are now resistant to T1 infection. In order to explain this resistance, two hypotheses were suggested. The **adaptation hypothesis** was that some bacteria became immune to T1 upon exposure to the bacteriophage. The key idea in this hypothesis is that the exposure is necessary for the resistance to develop. Alternatively, the bacteria may become immune through **spontaneous mutation.**

In order to distinguish between these hypothesis, Luria and Delbrück (1943) grew small individual cultures of T1-sensitive *E. coli*. From each culture equal quantities were added to several Petri dishes containing T1. Each dish was scored for the number of surviving (hence phage-resistant) colonies.

If the adaptation hypothesis is correct, each bacterium would have a small probability of developing resistance after being exposed to T1. Hence, using the Poisson limit to the binomial distribution, we would expect all the dishes to show Poisson variability (variance equal to the mean), regardless of which culture they came from. If, on the other hand, spontaneous mutation occurs at a constant rate, we would expect large numbers of resistant colonies originating from cultures in which mutation took place early in the experiment, and would therefore expect super-Poisson variability (variance larger than the mean) between dishes from different cultures.

In order to develop a control group, Luria and Delbrück also computed numbers of mutations in dishes that were all originating in one large culture. Since a large culture is constantly mixed, we would expect Poisson variability in these dishes, regardless of which hypothesis is true. Table 1.1 contains some results.

Table 1.1 Mutation numbers

Dish	Same culture	Different cultures
1	14	6
2	15	5
3	13	10
4	21	8
5	15	24
6	14	13
7	26	165
8	16	15
9	20	6
10	13	10

The sample mean for the control group was 16.7, with a sample variance of 15.0, while the experimental group had a sample mean of 26.2 with a sample variance of 2178. The difference between the means is not statistically significant, as assessed by a two-sample t-test. From comparing the variances it is, however, very clear that the mutation hypothesis is much better supported than the adaptation hypothesis, as has been verified by later experiments.

ese hypothesis, Luria and Delbrück
1-sensitive *E. coli*. From each culture
etri dishes containing T1. Each dish
nce phage-resistant) colonies.

t, each bacterium would have a small
being exposed to T1. Hence, using the
, we would expect all the dishes to
o the mean), regardless of which cul-
d, spontaneous mutation occurs at a
bers of resistant colonies originating
e early in the experiment, and would
y (variance larger than the mean)

, Luria and Delbrück also computed
e all originating in one large culture.
we would expect Poisson variability
hesis is true. Table 1.1 contains some

tion numbers

Different cultures
6
5
10
8
24
13
165
15
6
10

roup was 16.7, with a sample variance
d a sample mean of 26.2 with a sample
ween the means is not statistically
t-test. From comparing the variances it
n hypothesis is much better supported
en verified by later experiments.

Many natural scientists are of the opinion that most problems on a macroscopic scale can be solved using deterministic models, at least in principle. Modern chaos theory suggests that some of these models are nonlinear. Techniques of applied probability are often thought of as what may be called **statistical** models, in the sense of being reasonably approximate descriptions of the phenomenon under study, while lacking much physical basis. Our point of view will be somewhat different. There are frequently reasons for using stochastic models (as opposed to deterministic ones) to *model*, rather than just describe, various natural phenomena. A distinction is sometimes made between **forecast** models and **structural** models. The former do not aim at describing the phenomenon under study, but rather try to predict outcomes, while the latter are more focused on describing the processes that produce the phenomenon. Our interest lies mainly in the latter. Some of the potential advantages in using stochastic models are listed below.

Aids understanding of the phenomenon studied. We will see (section 3.8) an example from neurophysiology where a stochastic model can rule out several proposed mechanisms for communication between nerve cells. While it is not possible to observe directly the underlying kinetics, it is possible to ascertain whether or not a certain neurological channel is open or closed. The time spent open can be used to estimate, based on simple assumptions, the actual kinetics of the channel.

More versatile than deterministic models. An example from entomology illustrates this. In a long series of experiments, the Australian entomologist A. J. Nicholson observed the population dynamics of blowflies, an insect that lays its eggs in the skin of sheep. The larvae feed off the flesh of the sheep, and can kill it if the infestation is severe enough. Nicholson's experiments used varying regimes of food, population density, and initial population. The data consist of observations only on the number of flies that were born and died, respectively, on a given day (Exercise 2.D6 in the next chapter gives one such data set). Guckenheimer et al. (1976) developed a model of the form

$$\frac{\partial n(t,a)}{\partial t} + \frac{\partial n(t,a)}{\partial a} = -d(a,n_t)n(t,a), \tag{1.24}$$

where $n(t,a)$ is the number of flies aged a at time t, while n_t is the total population size at time t. This model, even for fairly simple choices of death rate d, can exhibit chaotic behavior. However, since the age distribution of the population was not observed, this model could not directly be fitted to the data, and arbitrary assumptions about the age distribution could only be investigated by the qualitative behavior of numerical solutions of the partial differential equation (1.24). On the other hand, using a stochastic model, Brillinger et al. (1980) were able to reconstruct (with a specified amount of uncertainty) the age distribution of the population, and deduce that the population dynamics was both age and density dependent. In addition, one could infer from the stochastic model that the physiology of the population was changing by natural selection over the

course of the experiment (Guttorp, 1980).

Allows assessment of variability. When modeling the long range transport of pollutants in the atmosphere, one important problem is to identify the source from observations at the point of deposition. There are both simple and complicated deterministic models in use that attempt to perform this task (Guttorp, 1986). Such models must take into account differing emission sources, transportation paths, chemical transformations in the atmosphere, and different types of deposition. The output of these models can be used to allocate sources as a fraction of total deposition. While such numbers give an impression of objective scientific analysis, it is extremely difficult to establish their uncertainty. Such uncertainty results from uncertainty in emissions data, model errors in describing the complicated processes involved, measurement errors in the field, laboratory errors, etc. (Guttorp and Walden, 1987). If, on the other hand, a carefully built stochastic model were used (Grandell, 1985, is a general introduction to such models) it would be possible to set confidence bands for source apportionment.

Extension of deterministic models. We saw in section 1.1 how a simple deterministic model of the motion of a coin can be combined with a stochastic model of the initial conditions to yield a probabilistic description of the outcome of coin tosses. It is a lot easier to test the combined model than to test the purely deterministic model, in which the initial conditions must be exactly realized.

Data compression. In the early days of mathematical statistics there was a great deal of emphasis on obtaining the maximum amount of information out of a given, often small, set of data. Many current data-gathering techniques yield vast data sets, which need to be described compactly. The emphasis is on compression while losing the minimum amount of information. For example, such techniques are needed to analyze satellite images (Baddeley et al., 1991). This area, compression of data sets using parametric models describing images, is likely to become one of the most important areas of statistical research over the next few decades in view of the rapid increase in remote sensing applications.

1.4. Overview

In Chapter 2 we introduce Markov chains with discrete time and discrete state space. The concept of reversibility of Markov chains has important consequences, and can sometimes be related to physical considerations of detailed balance in closed systems. Results on classification of states are needed to see under what circumstances statistical inference is possible. We work through the asymptotics of transition probabilities in some detail, partly to establish some necessary foundations for Markov chain Monte Carlo methods that will be very important in Chapter 4. We do both nonparametric and parametric statistical theory, present a linear model for higher order Markov chains, and finally look at hidden Markov processes and their reconstruction from noisy data. This is a

modeling the long range transport of
ant problem is to identify the source
n. There are both simple and compli-
ttempt to perform this task (Guttorp,
t differing emission sources, transpor-
the atmosphere, and different types of
can be used to allocate sources as a
umbers give an impression of objec-
ifficult to establish their uncertainty.
y in emissions data, model errors in
lved, measurement errors in the field,
n, 1987). If, on the other hand, a care-
Grandell, 1985, is a general introduc-
le to set confidence bands for source

saw in section 1.1 how a simple deter-
be combined with a stochastic model
ilistic description of the outcome of
mbined model than to test the purely
onditions must be exactly realized.

f mathematical statistics there was a
aximum amount of information out of
rrent data-gathering techniques yield
ibed compactly. The emphasis is on
amount of information. For example,
tellite images (Baddeley et al., 1991).
parametric models describing images,
rtant areas of statistical research over
d increase in remote sensing applica-

s with discrete time and discrete state
Markov chains has important conse-
to physical considerations of detailed
assification of states are needed to see
ence is possible. We work through the
some detail, partly to establish some
Monte Carlo methods that will be very
nparametric and parametric statistical
order Markov chains, and finally look
construction from noisy data. This is a

first (and very simple) stab at some Bayesian techniques that will be very useful in image reconstruction later.

Chapter 3 switches to continuous time, while still retaining the Markov property. Many of the results from the previous chapter can be used on appropriate discrete time subchains to establish properties of the continuous time models. As in the case of discrete time, we do both parametric and nonparametric inference, and study partially observed as well as completely hidden Markov models.

In the fourth chapter we move to random fields. We look at nearest neighbor interaction potentials and relate them to the Markov property. We encounter the phenomenon of phase transition in a simple example. In this chapter we use parametric statistical inference, making heavy use of Markov chain Monte Carlo techniques. Hidden processes are reconstructed using the Bayesian approach first encountered at the end of Chapter 2, and a general formulation of Markov random fields on graphs is found to be useful.

In the previous chapters, non-Markovian processes were introduced as noisy functions of an underlying (but perhaps not observable) Markovian structure. In Chapter 5 we turn to a class of mostly non-Markovian processes which do not have any underlying Markovian structure, but are developing dynamically depending on their entire previous history. These processes model events that occur separately over time. A variety of ways of thinking about such point processes are discussed, and we encounter both parametric and nonparametric statistical inference.

The Brownian motion process is the foundation of continuous time Markov processes with continuous paths. Such processes are discussed in Chapter 6. We illustrate how one can build diffusion processes using stochastic differential equations. The statistical inference becomes quite complicated here, since it is no longer straightforward to define a likelihood.

1.5. Bibliographic remarks

Much of the material in section 1.1 follows Keller (1986) and Engel (1992). An elementary introduction to the topic of chaotic dynamics is Ruelle (1991). More general discussion of chaos, indeterminism, and randomness is in a report from the Royal Statistical Society Meeting on Chaos (1992).

The importance of the Brook expansion was made clear by Besag (1974). Any measure-theoretic introduction to probability (such as Billingsley, 1979, section 36) explains the use of the Kolmogorov consistency condition.

The *E. coli* example follows Klug and Cummings (1991), pp. 409–412.

1.6. Exercises

Theoretical exercises

1. Suppose that q satisfies the positivity condition (1.14). Show that if $\mathbf{x}$ and $\mathbf{y}$ are outcomes such that $q(\mathbf{x})>0$ and $q(\mathbf{y})>0$, then $q(\mathbf{x}_1^{i-1},\mathbf{y}_i^m)>0$.

2. For $0\le t_1<\cdots<t_n$ and $0\le r_1\le\cdots\le r_n$, define a stochastic process $X(t)$ by assuming that

$$\mathbf{P}(X(t_1)=r_1,\ldots,X(t_n)=r_n) = \frac{\lambda^{r_n}e^{-\lambda t_n}t_1^{r_1}(t_2-t_1)^{r_2-r_1}\cdots(t_n-t_{n-1})^{r_n-r_{n-1}}}{r_1!(r_2-r_1)!\cdots(r_n-r_{n-1})!}$$

Show that this process satisfies the Kolmogorov consistency condition.

3. For the process in Exercise 2, write $X_i=X(t_i)$. Show that $q(\mathbf{x})=\mathbf{P}(\mathbf{X}=\mathbf{x})$ satisfies the positivity condition.

4. Consider a (theoretical) roulette wheel divided into n sections that are alternating red and black. Assume that the arcs are all of equal length. The wheel is spun using a large initial impulse, and the angle X in radians (with respect to a fixed point outside the wheel) of a given spot on the wheel (at one end of a black section) is assumed to have a density $f(x)$.

(a) Show that the probability of black, i.e., that the fixed point is nearest a black section, can be computed as

$$\mathbf{P}(\text{black}) = \mathbf{P}(\frac{nX}{4\pi} \bmod(1) \le \tfrac{1}{2}).$$

(b) Poincaré (1896) showed that under mild regularity conditions on the density of X one has that tX mod(1) converges in distribution to a uniform random variable on (0,1) as $t\to\infty$. Apply this result to deduce the limiting probability of black as $n\to\infty$.

Remark: The method used here uses no physics, only the existence of a sufficiently regular density. It is called the method of **arbitrary functions** (see Engel, 1992, for some historical background).

Computing exercises

C1. Simulate the process described in Exercise 4(a), and assess the convergence of $\mathbf{P}(\text{black})$ to ½ as a function of the number n of sectors. Modify the process so that the red sections have arc length R and the black have arc length B. How does that affect the probability? The convergence?

ondition (1.14). Show that if x and y
, then $q(\mathbf{x}_1^{i-1},\mathbf{y}_i^m)>0$.
define a stochastic process $X(t)$ by

$$\frac{(t_2-t_1)^{r_2-r_1}\cdots(t_n-t_{n-1})^{r_n-r_{n-1}}}{(r_1)!\cdots(r_n-r_{n-1})!}$$

gorov consistency condition.
$X_i=X(t_i)$. Show that $q(\mathbf{x})=\mathbf{P}(\mathbf{X}=\mathbf{x})$

divided into n sections that are alter-
s are all of equal length. The wheel is
e angle X in radians (with respect to a
n spot on the wheel (at one end of a
y $f(x)$.
., that the fixed point is nearest a black

½).

ild regularity conditions on the density
distribution to a uniform random vari-
to deduce the limiting probability of

no physics, only the existence of a
ne method of **arbitrary functions** (see
und).

ercise 4(a), and assess the convergence
ber n of sectors. Modify the process so
and the black have arc length B. How
vergence?

Data exercises

D1. Suppose we are observing a roulette wheel of the type described in Exercise 4 (so it has no zero, or, equivalently, we ignore the zeroes) with $n=32$.

(a) Derive the distribution of runs, i.e., successive strings of the same color, assuming that the probability of black is ½.

(b) Pearson (1900) reported data on runs from the casino in Monte Carlo in July of 1892. The data are given in Table 1.2.

Table 1.2 Runs of the same color in Monte Carlo

Run length	1	2	3	4	5	6
Count	2462	945	333	220	135	81
Run length	7	8	9	10	11	12
Count	43	30	12	7	5	1

Do they agree with the distribution in (a)? If not, can you explain the discrepancy?