Statistical Analysis of Stochastic Processes in Time

Many observed phenomena, from the changing health of a patient to values on the stock market, are characterised by quantities that vary over time: stochastic processes are designed to study them. Much theoretical work has been done but virtually no modern books are available to show how the results can be applied. This book fills that gap by introducing practical methods of applying stochastic processes to an audience knowledgeable only in the basics of statistics. It covers almost all aspects of the subject and presents the theory in an easily accessible form that is highlighted by application to many examples. These examples arise from dozens of areas, from sociology through medicine to engineering. Complementing these are exercise sets making the book suited for introductory courses in stochastic processes.

Software is provided within the freely available R system for the reader to be able to apply all the models presented.

J. K. LINDSEY is Professor of Quantitative Methodology, University of Liège. He is the author of 14 books and more than 120 scientific papers.

CAMBRIDGE SERIES IN STATISTICAL AND PROBABILISTIC MATHEMATICS

This series of high-quality upper-division textbooks and expository monographs covers all aspects of stochastic applicable mathematics. The topics range from pure and applied statistics to probability theory, operations research, optimization, and mathematical programming. The books contain clear presentations of new developments in the field and also of the state of the art in classical methods. While emphasizing rigorous treatment of theoretical methods, the books also contain applications and discussions of new techniques made possible by advances in computational practice.

Already published

1. *Bootstrap Methods and Their Application*, by A. C. Davison and D. V. Hinkley
2. *Markov Chains*, by J. Norris
3. *Asymptotic Statistics*, by A. W. van der Vaart
4. *Wavelet Methods for Time Series Analysis*, by Donald B. Percival and Andrew T. Walden
5. *Bayesian Methods*, by Thomas Leonard and John S. J. Hsu
6. *Empirical Processes in M-Estimation*, by Sara van de Geer
7. *Numerical Methods of Statistics*, by John F. Monahan
8. *A User's Guide to Measure Theoretic Probability*, by David Pollard
9. *The Estimation and Tracking of Frequency*, by B. G. Quinn and E. J. Hannan
10. *Data Analysis and Graphics using R*, by John Maindonald and John Braun
11. *Statistical Models*, by A. C. Davison
12. *Semiparametric Regression*, by David Ruppert, M. P. Wand and R. J. Carroll
13. *Exercises in Probability*, by L. Chaumont and M. Yor

Statistical Analysis of Stochastic Processes in Time

J. K. Lindsey
University of Liège

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town,
Singapore, São Paulo, Delhi, Mexico City

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521837415

First published 2004

A catalogue record for this publication is available from the British Library

ISBN 978-0-521-83741-5 Hardback

Contents

Preface

Throughout their history, human beings have been fascinated by time. Indeed, what is history but an interpretation of time? Each civilisation has had its own special conception of time. Our present anti-civilisation only knows 'time is money'! No one can deny that the study of time is important. This text attempts to make more widely available some of the tools useful in such studies.

Thus, my aim in writing this text is to introduce research workers and students to ways of modelling a wide variety of phenomena that occur over time. My goal is explicitly to show the broadness of the field and the many inter-relations within it. The material covered should enable mathematically literate scientists to find appropriate ways to handle the analysis of their own specific research problems. It should also be suitable for an introductory course on the applications of stochastic processes. It will allow the instructor to demonstrate the unity of a wide variety of procedures in statistics, including connections to other courses. If time is limited, it will be possible to select only certain chapters for presentation.

No previous knowledge of stochastic processes is required. However, an introductory course on statistical modelling, at the level of Lindsey (2004), is a necessary prerequisite. Although not indispensable, it may be helpful to have more extensive knowledge of several areas of statistics, such as generalised linear and categorical response models. Familiarity with classical introductory statistics courses based on point estimation, hypothesis testing, confidence intervals, least squares methods, personal probabilities, . . . will be a definite handicap.

Many different types of stochastic processes have been proposed in the literature. Some involve very complex and intractable distributional assumptions. Here, I shall restrict attention to a selection of the simpler processes, those for which explicit probability models, and hence likelihood functions, can be specified and which are most useful in statistical applications modelling empirical data. More complex models, including those requiring special estimation techniques such as Monte Carlo Markov Chain, are beyond the scope of this text. Only parametric models are covered, although descriptive 'nonparametric' procedures, such as the Kaplan–Meier estimates, are used for examining model fit.

The availability of explicit probability models is important for at least two reasons:

(i) Probability statements can be made about observable data, including the observed data:
   (a) A likelihood is available for making inferences.
   (b) Predictions can be made.
(ii) If the likelihood can be calculated, models can be compared to see which best fit the data, instead of making empty claims about wonderful models with no empirical basis, as is most often done in the statistical literature.

Isolated from a probability model basis, parameter estimates, with their standard errors, are of little scientific value.

Many standard models, such as those for survival, point processes, Markov chains, and time series, are presented. However, because of the book's wide scope, it naturally cannot cover them in as great a depth as a book dedicated to only one of them. In addition, certain areas, such as survival analysis and time series, occupy vast literatures to which complete justice cannot be made here. Thus, in order to provide a reasonably equitable coverage, these two topics are explored especially briefly; the reader can consult a good introductory text on either of these topics for additional details.

Many basic theoretical results are presented without proof. The interested reader can pursue these in further detail by following up the 'Further reading' list at the end of each chapter. On the other hand, for the readers primarily interested in developing appropriate stochastic models to apply to their data, the sections labelled 'Theory' can generally be skimmed or skipped and simply used as a reference source when required for deeper understanding of their applications.

Stochastic processes usually are classified by the type of recording made, that is, whether they are discrete events or continuous measurements, and by the time spacing of recording, that is, whether time is discrete or continuous. Applied statisticians and research workers usually are particularly interested in the type of response so that I have chosen the major division of the book in this way, distinguishing between categorical events and continuous measurements. Certain models, such as Markov chains using logistic or log linear models, are limited to discrete time, but most of the models can be applied in either discrete or continuous time.

Classically, statistics has distinguished between linear and nonlinear models, primarily for practical reasons linked with numerical methods and with inference procedures. With modern computing power, such a distinction is no longer necessary and will be ignored here. The main remaining practical difference is that nonlinear models generally require initial values of parameters to be supplied in the estimation procedure, whereas linear models do not.

It is surprisingly difficult to find material on fitting stochastic models to data. Most of the literature concentrates either on the behaviour of stochastic models under specific restrictive conditions, with illustrative applications rarely involving real data, or on the estimation of some asymptotic statistics, such as means or variances. Unavoidably, most of the references for further reading given at the ends of the chapters are of much more difficult level than the present text.

My final year undergraduate social science students have helped greatly in developing this course over the past 25 years. The early versions of the course were based on Bartholomew (1973), but, at that time, it was very difficult or impossible actually to analyse data in class using those methods. However, this rapidly evolved, eventually to yield Lindsey (1992). The present text reflects primarily the more powerful software now available. Here, I have supplemented the contents of my current course with extra theoretical explanations of the stochastic processes and with examples drawn from a wide variety of areas besides the social sciences.

Thus, I provide the analysis of examples from many areas, including botany (leaf growth), criminology (recidivism), demography (migration, human mortality), economics and finance (capital formation, share returns), education (university enrolment), engineering (degradation tests, road traffic), epidemiology (AIDS cases, respiratory mortality, spermarche), industry (mining accidents), medicine (blood pressure, leukæmia, bladder and breast cancer), meteorology (precipitation), pharmacokinetics (drug efficacy, radioactive tracers), political science (voting behaviour), psychology (animal learning), sociology (divorces, social mobility), veterinary science (cow hormones, sheep infections), and zoology (locust activity, nematode control). Still further areas of application are covered in the exercises.

The data for the examples and exercises, as well as the R code for all of the examples, can be found at `popgen0146uns50.unimaas.nl/~jlindsey`, along with the required R libraries. With this material, the reader can see exactly how I performed the analyses described in the text and adapt the code to his or her particular problems.

This text is not addressed to probabilists and academic statisticians, who will find the definitions unrigorous and the proofs missing. Rather, it is aimed at the scientist, looking for realistic statistical models to help in understanding and explaining the specific conditions of his or her empirical data. As mentioned above, the reader primarily interested in applying stochastic processes can omit reading the theory sections and concentrate on the examples. When necessary, reference can then be made to the appropriate parts of theory.

I thank Bruno Genicot, Patrick Lindsey, and Pablo Verde who provided useful comments on earlier versions of certain chapters of this text.

Notation and symbols

Notation is generally explained when it is first introduced. However, for reference, some of the more frequently used symbols are listed below.

Vectors are bold lower case and matrices bold upper case Greek or Roman letters. $\top$ denotes the transpose of a vector or matrix.

h, i, j, k, l	arbitrary indices
Y, y	random response variable and its observed value
T, t	time
h	lag
S	sum of random variables
x, $\mathbf{X}$	explanatory variables
N, n	number of events
ΔN, Δn	change in number of events
Δt	interval width
$\mathcal{F}$	previous history
μ	location parameter
σ^2	variance
π	probability (usually binary)
τ	change point parameter
ψ, ξ	random parameters
α, β, δ, ϵ, ζ, ϕ, ν, υ, κ, θ	arbitrary parameters
ρ	(auto)correlation or other dependence parameter
M	order of a Markov process
R	length of a series
$\Pr(y)$	probability of response
$f(y)$	probability density function
$F(y)$	cumulative distribution function
$S(y)$	survival function
$p(\cdot)$	probability of a random parameter
$h(\cdot)$	arbitrary regression function
$g(\cdot)$	link function
Λ	integrated intensity (function)
λ	intensity (function)

$\mathrm{E}(\cdot)$	expected value
γ	(auto)covariance (function)
$\mathbf{\Sigma}$	covariance matrix
$\mathrm{I}(\cdot)$	indicator function
$B(\cdot,\cdot)$	beta function
$\Gamma(\cdot)$	gamma function
$\mathrm{L}(\cdot)$	likelihood function
$\mathbf{T}$	transition (probability) matrix
$\mathbf{\Lambda}$	transition intensity matrix
$\boldsymbol{\pi}$	marginal or conditional probability distribution
$\boldsymbol{\nu}$	first passage distribution
$\mathbf{F}$	diagonal matrix of probabilities
$\mathbf{D}$	diagonal matrix of eigenvalues
$\mathbf{W}$	matrix of eigenvectors
$\mathbf{b}$	vector of deterministic input
$\boldsymbol{\epsilon}$	vector of random input