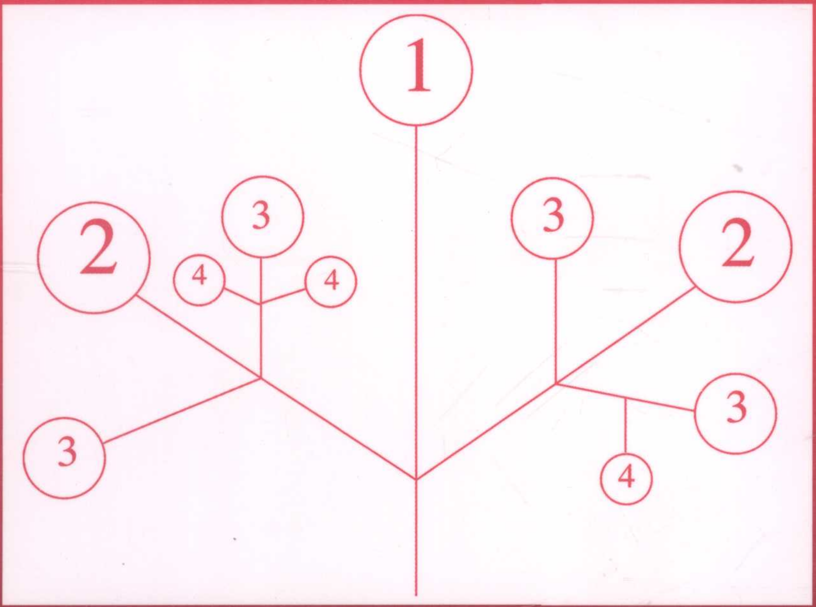
Texts in Statistical Science

Applied Stochastic Modelling

Second Edition



Byron J. T. Morgan

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Texts in Statistical Science

Applied Stochastic Modelling

Second Edition

Byron J. T. Morgan
University of Kent
UK

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

Chapman & Hall/CRC
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

No claim to original U.S. Government works
Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-13: 978-1-58488-666-2 (Softcover)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Applied Stochastic Modelling

Second Edition

CHAPMAN & HALL/CRC
Texts in Statistical Science Series

Analysis of Failure and Survival Data
P. J. Smith

The Analysis of Time Series— An Introduction, Sixth Edition
C. Chatfield

Applied Bayesian Forecasting and Time Series Analysis
A. Pole, M. West and J. Harrison

Applied Nonparametric Statistical Methods, Fourth Edition
P. Sprent and N.C. Smeeton

Applied Statistics — Handbook of GENSTAT Analysis
E.J. Snell and H. Simpson

Applied Statistics — Principles and Examples
D.R. Cox and E.J. Snell

Applied Stochastic Modelling, Second Edition
B.J.T. Morgan

Bayesian Data Analysis, Second Edition
A. Gelman, J.B. Carlin, H.S. Stern and D.B. Rubin

Bayesian Methods for Data Analysis, Third Edition
B.P. Carlin and T.A. Louis

Beyond ANOVA — Basics of Applied Statistics
R.G. Miller, Jr.

Computer-Aided Multivariate Analysis, Fourth Edition
A.A. Afifi and V.A. Clark

A Course in Categorical Data Analysis
T. Leonard

A Course in Large Sample Theory
T.S. Ferguson

Data Driven Statistical Methods
P. Sprent

Decision Analysis — A Bayesian Approach
J.Q. Smith

Elementary Applications of Probability Theory, Second Edition
H.C. Tuckwell

Elements of Simulation
B.J.T. Morgan

Epidemiology — Study Design and Data Analysis, Second Edition
M. Woodward

Essential Statistics, Fourth Edition
D.A.G. Rees

Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models
J.J. Faraway

A First Course in Linear Model Theory
N. Ravishanker and D.K. Dey

Generalized Additive Models: An Introduction with R
S. Wood

Interpreting Data — A First Course in Statistics
A.J.B. Anderson

An Introduction to Generalized Linear Models, Third Edition
A.J. Dobson and A.G. Barnett

Introduction to Multivariate Analysis
C. Chatfield and A.J. Collins

Introduction to Optimization Methods and Their Applications in Statistics
B.S. Everitt

Introduction to Probability with R
K. Baclawski

Introduction to Randomized Controlled Clinical Trials, Second Edition
J.N.S. Matthews

Introduction to Statistical Methods for Clinical Trials
T.D. Cook and D.L. DeMets

Large Sample Methods in Statistics
P.K. Sen and J. da Motta Singer

Linear Models with R
J.J. Faraway

Markov Chain Monte Carlo — Stochastic Simulation for Bayesian Inference, Second Edition
D. Gamerman and H.F. Lopes

Mathematical Statistics
K. Knight

Modeling and Analysis of Stochastic Systems
V. Kulkarni

Modelling Binary Data, Second Edition
D. Collett

Modelling Survival Data in Medical Research, Second Edition
D. Collett

Multivariate Analysis of Variance and Repeated Measures — A Practical Approach for Behavioural Scientists
D.J. Hand and C.C. Taylor

Multivariate Statistics — A Practical Approach
B. Flury and H. Riedwyl

Pólya Urn Models
H. Mahmoud

Practical Data Analysis for Designed Experiments
B.S. Yandell

Practical Longitudinal Data Analysis
D.J. Hand and M. Crowder

Practical Statistics for Medical Research
D.G. Altman

A Primer on Linear Models
J.F. Monahan

Probability — Methods and Measurement
A. O'Hagan

Problem Solving — A Statistician's Guide, Second Edition
C. Chatfield

Randomization, Bootstrap and Monte Carlo Methods in Biology, Third Edition
B.F.J. Manly

Readings in Decision Analysis
S. French

Sampling Methodologies with Applications
P.S.R.S. Rao

Statistical Analysis of Reliability Data
M.J. Crowder, A.C. Kimber,
T.J. Sweeting, and R.L. Smith

Statistical Methods for Spatial Data Analysis
O. Schabenberger and C.A. Gotway

Statistical Methods for SPC and TQM
D. Bissell

Statistical Methods in Agriculture and Experimental Biology, Second Edition
R. Mead, R.N. Curnow, and A.M. Hasted

Statistical Process Control — Theory and Practice, Third Edition
G.B. Wetherill and D.W. Brown

Statistical Theory, Fourth Edition
B.W. Lindgren

Statistics for Accountants
S. Letchford

Statistics for Epidemiology
N.P. Jewell

Statistics for Technology — A Course in Applied Statistics, Third Edition
C. Chatfield

Statistics in Engineering — A Practical Approach
A.V. Metcalfe

Statistics in Research and Development, Second Edition
R. Caulcutt

Survival Analysis Using S — Analysis of Time-to-Event Data
M. Tableman and J.S. Kim

The Theory of Linear Models
B. Jørgensen

Time Series Analysis
H. Madsen

Preface to the Second Edition

The structure of the second edition of this book is very similar to that of the first edition; however, there have been numerous changes throughout. In particular, a large number of new exercises have been added, there is a new appendix on computational methods, and the discussion of Bayesian methods has been extended. The bibliography has been updated, throughout figures have been improved, and, where necessary, errors have been corrected. I am grateful for the many positive comments that the first version of the book received, and to those who have written to point out mistakes, and ways in which the book might be improved. I thank especially Ted Catchpole, Rachel Fewster, Ruth King, Rachel McCrea, David Miller, Karen Palmer and Martin Ridout. I thank MATLAB® for providing the latest version of the package, and I am also grateful for the help and patience of Rob Calver and colleagues at CRC Chapman & Hall. The book continues to be the set text for final year under-graduate and post-graduate courses in respectively Applied Stochastic Modelling and Data Analysis, and Computational Statistics in the University of Kent. The book successfully revises and integrates the probability and statistics methods of earlier lecture courses. At the same time it brings students into contact with modern computational methods; it provides students with practical experience of scientific computing at use in applied statistics, in the context of a range of interesting real-life applications. The book Web Site contains the data sets from the book, the MATLAB computer programs, as well as corresponding versions in R. There is a solutions manual for the exercises and the computer practical sheets that are used in Kent. The book has been the basis for a course on *Applied Stochastic Modelling* held at Pfizer Central Research in Sandwich, Kent, during 2007, and also for a Continuing Professional Development course with the same name, to be held at the Royal Statistical Society, London, in 2008. The slides for these courses are also available on the book Web site.

Canterbury

Preface

This book is being completed at the end of a millenium, the last 30 years of which have seen many exciting major developments in the theory and practice of statistics. The book presents many of the most important of these advances. It has its origins in a series of talks on aspects of modern statistics for fitting stochastic models to data, commissioned by statisticians at Pfizer Central Research in Sandwich in Kent. These talks gave rise to a 30-hour lecture course given to third year undergraduates, statistics MSc students and first-year statistics PhD students at the University of Kent, and this book has grown from the notes for that lecture course. These students have found that even the most recently developed statistical methods may be readily understood and successfully applied in practice. As well as covering modern techniques, the material of the book integrates and revises standard probability and statistical theory. Much modern statistical work is implemented using a computer. Thus it is necessary in a book of this nature to include computer instructions of some kind. The integrated computer language MATLAB has been selected for this purpose, and over 50 short MATLAB programs are included throughout the book. Their distribution and purpose are described in the index of MATLAB programs. They may all be accessed from the book site on the World Wide Web. They are designed to be illustrative, rather than completely efficient. Often doubts concerning theory are dissipated when one can see computer code for the theory. Students of the book material have certainly found the MATLAB programs to be a useful aid to learning. It has been uplifting to observe the satisfaction that students have gained from running the MATLAB programs of the book. Often the students had no previous knowledge of MATLAB and also little prior experience of scientific computing. The material of Appendix B, which summarises important features of MATLAB, and tutorial assistance were all that was needed by these students. However it should be stressed that while the computer programs are included as an aid to learning, the book may be read and used without reference to the programs. S-plus versions of the programs are available on the book site on the World Wide Web. There are also some references to the use of symbolic algebra packages such as MAPLE, as these provide powerful tools for stochastic modelling.

The target student audience for the book is final-year undergraduate and MSc students of mathematics and statistics. The book is also intended as a single convenient source of reference for research scientists and post-graduate students, using modern statistical methods which are currently described in depth in a range of single-topic textbooks. Prior knowledge is assumed at the level of a typical second-year university course on probability and statistics. Appendix A summarises a number of important formulae and results from probability and statistics. A small fraction of the book sections and exercises contain more advanced material, and these are starred. Kernel density estimation is a central aspect of modern statistics, and therefore the basic ideas are summarised in Appendix C. While a limited number of exercises have solutions included in the book, a more extensive set of solutions is to be found on the World Wide Web book site.

Statistical methods are all-pervasive, contributing significantly to subjects as diverse as geology, sociology, biology and economics. The construction, fitting and evaluation of statistical and stochastic models are not only vitally important in areas such as these, but they are also great fun. It is hoped that some of the enjoyment and fascination of the subject will be gained by readers of this book.

The book is motivated by real data and problems. The examples and exercises are often chosen from my own experience and, as can be seen from the index of data sets, many have arisen from biology. The areas covered are sometimes atypical, and sometimes classical, such as survival analysis, quantal assay and capture-recapture. Several of the examples recur at various points throughout the book. The data are available on the book site on the World Wide Web.

Acknowledgments

I owe a great debt of gratitude to the many scientists and statisticians with whom I have collaborated, and with whom I continue to collaborate. I am especially grateful to my MATLAB mentor, Ted Catchpole, who with Paul Terrill, contributed some of the MATLAB programs. The original proposal for a set of lectures at Pfizer Central Research was agreed with Trevor Lewis, Group Executive Director, Biometrics, at Pfizer. The Pfizer audience was typically challenging, with Ken Martin and John Parrott in particular regularly asking probing questions. Steve Brooks, Steve Buckland, Martin Ridout, and Paul Terrill all commented helpfully on draft chapters. At Arnold, Nicki Dennis, and Kirsty Stroud were patient and encouraging. Three years of Kent students provided a sympathetic test bed and launching pad. The many drafts of the book were typed enthusiastically by Lilian Bond, Julie Snook, and Mavis Swain.

Canterbury

‘For the things we have to learn before we can do them, we learn by doing them’

Aristotle

Contents

CHAPTER 1

Introduction and Examples

1.1 Introduction

Reported falls in human sperm counts in many developed countries have serious implications for the future of mankind. In *fecundability* studies, data are collected on waiting times to conception in human beings, as well as on variables such as age and body mass index, which is a measure of obesity. For instance, the paper by Jensen et al. (1998) concluded that the probability of conception in a menstrual cycle was lowered if only five alcoholic drinks were taken by the woman each week. Data from studies such as this require appropriate statistical analysis, which quite often results from describing the data by means of models tailored specifically to the particular question of interest.

In this book we shall consider a wide range of data sets. In each case our objective is to find a succinct description of the data, which may then be used either as a summary or to provide the basis for comparison with other examples. We shall do this by proposing simple models, and then fitting them to the data. The models we shall consider are based on the axioms of probability theory, as described, for example, by Grimmett and Stirzaker (1992). A number of useful results and formulae are to be found in Appendix A. Our models may be described in various ways. We may describe them as *probability* models, when the emphasis is on the probabilistic formulation, or we may describe the models as *statistical*, when there is a greater emphasis on fitting the models to data. Many of the models we describe in this book develop over time or space, and they are then called *stochastic*. What is involved in statistical modelling is readily appreciated from considering a fecundability example.

Example 1.1: Fecundability

Table 1.1 describes the number of fertility cycles to conception required by fertile human couples setting out to conceive. The data were collected retrospectively, which means that information was only obtained from women who had conceived, and the women involved have been classified according to whether they smoked or not. Couples requiring more than 12 cycles are grouped together in a single category.

The couples in this study are essentially waiting for an event, and the simplest probability model for waiting times when they are integers, as here, is the geometric model. Let X denote the number of cycles to conception, and

Table 1.1 *Cycles to conception, classified by whether the female of the couple smoked or not. The data, taken from Weinberg and Gladen (1986), form a subset of data presented by Baird and Wilcox (1985). Excluded were women whose most recent method of contraception was the pill, as prior pill usage is believed to reduce fecundability temporarily. The definition of "smoking" is given in the source papers.*

Cycle	Women smokers	Women non-smokers
1	29	198
2	16	107
3	17	55
4	4	38
5	3	18
6	9	22
7	4	7
8	5	9
9	1	5
10	1	3
11	1	6
12	3	6
>12	7	12
Total	100	486

let p be the probability of conception per cycle. If we assume p is constant, then

$$pr(X = k) = (1 - p)^{k-1}p, \quad \text{for } k \geq 1.$$

Noting that $pr(X = 1) = p$ allows us to fit this model to each column of data very simply, to give the following estimates of p, which we denote by $\tilde{p}$:

	$\tilde{p}$
Smokers	0.29
Non-smokers	0.41

On the basis of these results we then have the immediate conclusion that each cycle non-smokers are 41% more likely to conceive than smokers (since $41 \approx (41 - 29)/29$). The usefulness of this model is therefore apparent in providing a framework for simply comparing smokers and non-smokers. However, there are evident shortcomings to the model. For example, can we legitimately suppose that all couples behave the same way? Of course we cannot, and simply checking the fit of this model to the data will reveal its inadequacies. We shall consider later how we can make the model more elaborate to improve its description of the data. □

Standard terminology is that p in the above example is a *parameter*, and in fitting the model to the data we are producing the *estimator*, $\tilde{p}$. For this example, the parameter summarises the fecundability of a particular population. By itself, $\tilde{p}$ tells only part of the story, because it does not reflect the amount of data summarised by $\tilde{p}$. Naturally, because of the difference in sample sizes, the value of $\tilde{p}$ for non-smokers is more precise than that for smokers, and we can indicate the difference by constructing appropriate *confidence intervals*, or by associating with each estimator an estimate of its *standard error*, often just referred to as its *error*. We shall discuss ways of indicating precision, and present several examples, later in the book.

This book is devoted to the modelling process outlined in Example 1.1, but using a wide range of more sophisticated tools. We shall conclude this chapter with several more examples, which will be considered again later in the book.

1.2 Examples of data sets

Several of the models that we encounter in this book are general, with wide application, for instance multinomial models, models of survival and logistic regression models. Others have to be specially designed for particular problems and data sets, and we shall now consider a range of examples which require individually tailored models. However, the same basic statistical principles and approach are relevant in all cases. As we shall see, models for one application may also be of use in other areas which at first sight may appear to be quite different.

Example 1.2: Microbial infections

The data of Table 1.2 are taken from Sartwell (1950) and provide information on incubation periods of individuals who contracted streptococcal sore throat from drinking contaminated milk. Knowledge of when the contamination occurred allowed the incubation periods to be calculated.

Of interest here was fitting a stochastic model for the underlying behaviour of the infecting agents. Models of this kind are also used in the modelling of AIDS progression, in models of biological control and in models for the distribution of prions over cells. (Prions are the protein particles thought to result in mad cow disease.) □

Example 1.3: Polyspermy

Polyspermy occurs when two or more sperm enter an egg. It is relatively common in certain insects, fish, birds and reptiles. Fertilization of sea-urchin eggs has been of particular interest to biologists since the nineteenth century, and polyspermy in sea urchins *(Echinus esculentus)* has been examined by Rothschild and Swann (1950) and Presley and Baker (1970). An illustration of the experimental data that result is given in Table 1.3.

Thus for example, in the second of the four experiments, 84 eggs were exposed to sperm. After 15 seconds fertilisation was stopped by the addition of a spermicide, and it was found that 42 eggs had not been fertilised, 36 eggs

Table 1.2 *Sore throat incubation periods, given in units of 12 hrs.*

Incubation period	Number of individuals
0–1	0
1–2	1
2–3	7
3–4	11
4–5	11
5–6	7
6–7	5
7–8	4
8–9	2
9–10	2
10–11	0
11–12	1
>12	0

Table 1.3 *Polyspermy: distributions of sperm over eggs of sea urchins (data provided by Professor P.F. Baker).*

Experiment	No. of eggs in experiment	Length of experiment (secs)	Number of sperm in the egg					
			0	1	2	3	4	5
1	100	5	89	11	0	0	0	0
2	84	15	42	36	6	0	0	0
3	80	40	28	44	7	1	0	0
4	100	180	2	81	15	1	1	0

had been fertilised once and 6 eggs had been fertilised twice. The eggs in each experiment were different, and all four experiments took place under uniform conditions of temperature and sperm density.

Of interest here was a stochastic model incorporating the rates at which sperm enter the eggs. By means of the modelling it was possible to investigate how these rates vary over time. □

Example 1.4: Sprayed flour beetles

The data of Table 1.4 document the progress of flour beetles *Tribolium castaneum* sprayed with a well-known plant-based insecticide (pyrethrins B), and record the cumulative number of beetles that have died since the start of the experiment. The insecticide was sprayed at the given rates of application over small experimental areas in which the groups of beetles were confined