

**Statistical Models in**

**Epidemiology**

David Clayton

*Medical Research Council,  
Cambridge*

and

Michael Hills

*London School of Hygiene  
and Tropical Medicine*

OXFORD • NEW YORK • TOKYO  
OXFORD UNIVERSITY PRESS

Oxford University Press, Walton Street, Oxford OX2 6DP

Oxford New York Toronto  
Delhi Bombay Calcutta Madras Karachi  
Kuala Lumpur Singapore Hong Kong Tokyo  
Nairobi Dar es Salaam Cape Town  
Melbourne Auckland Madrid  
and associated companies in  
Berlin Ibadan

Oxford is a trade mark of Oxford University Press

Published in the United States  
by Oxford University Press Inc., New York

© David Clayton and Michael Hills, 1993

First published 1993  
Reprinted 1994 (twice)

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, without the prior permission in writing of Oxford University Press. Within the UK, exceptions are allowed in respect of any fair dealing for the purpose of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act, 1988, or in the case of reprographic reproduction in accordance with the terms of licences issued by the Copyright Licensing Agency. Enquiries concerning reproduction outside those terms and in other countries should be sent to the Rights Department, Oxford University Press, at the address above.

This book is sold subject to the condition that it shall not, by way of trade or otherwise, be lent, re-sold, hired out, or otherwise circulated without the publisher's prior consent in any form of binding or cover other than that in which it is published and without a similar condition including this condition being imposed on the subsequent purchaser.

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

Library of Congress Cataloging in Publication Data

Clayton, David, statistician.

Statistical models in epidemiology/David Clayton and Michael Hills.

Includes bibliographical references and index.

1. Epidemiology—Statistical methods. I. Hills, Michael. II. Title.  
[DNLM: 1. Epidemiology. 2. Models, Statistical. WA 105 C622s 1993]  
RA652.2M3C53 1993 614.4'072—dc20 93-19448

ISBN 0 19 852221 5 (Hbk)

Printed in Great Britain by  
Biddles Ltd, Guildford & King's Lynn

---

## Preface

---

The aim of this book is to give a self-contained account of the statistical basis of epidemiology. The book is intended primarily for students enrolled for a masters degree in epidemiology, clinical epidemiology, or biostatistics, and should be suitable both as the basis for a taught course and for private study.

Although we anticipate that most readers will have taken a first course in statistics, no previous knowledge is assumed, and the mathematical level of the book has been chosen to suit readers whose basic training is in biology. Some of the material in the book could be omitted at first reading, either because it is rather more demanding of mathematical skills or because it deals with rather specialized points. We have been careful to gather such material either into complete chapters or complete sections and to indicate these with a marginal symbol, as here.

Epidemiologists today have ready access to computer programs of great generality, but to use these sensibly and productively it is necessary to understand the ideas which lie behind them. The most important of these is the idea of a *probability model*. All statistical analysis of data is based on probability models, even though the models may not be explicit. Only by fully understanding the model can one fully understand the analysis.

Models depend on parameters, and values must be chosen for these parameters in order to match the model to the data. In showing how this is done we have chosen to emphasize the role of likelihood because this offers an approach to statistics which is both simple and intuitively satisfying. An additional advantage of this approach is that it requires the model and its parameters to be made explicit, even in the simplest situations. More complex problems can then be tackled by natural extensions of simple methods and do not require a whole new way of looking at things.

Most of the material in this book was developed during successive residential summer courses in epidemiology and statistics, held in Florence under the auspices of the European Educational Programme in Epidemiology. We are grateful to the International Agency for Cancer Research, the Regional Office for Europe of the World Health Organization, the Commission of the European Communities, and the Tuscany Regional Government, for sponsoring the program, and to Walter Davies, Organizing Secretary, and Rodolfo Saracci, Course Director, whose respective skills ensured that the course took place each year. We also acknowledge with thanks helpful

comments on earlier drafts from Damien Jolley, Bendix Carstensen, Dave Leon, and Nick Hills.

*Cambridge*  
*London*  
February 1993

David Clayton  
Michael Hills

### Dedication

To the students of the Florence course, 1988 – 92, without whose help and encouragement this book would never have appeared.

---

## Contents

---

I PROBABILITY MODELS AND LIKELIHOOD	
1	Probability models 3
2	Conditional probability models 10
3	Likelihood 18
4	Consecutive follow-up intervals 27
5	Rates 40
6	Time 53
7	Competing risks and selection 63 *
8	The Gaussian probability model 71
9	Approximate likelihoods 78
10	Likelihood, probability, and confidence 89
11	Null hypotheses and p-values 96
12	Small studies 110 *
13	Likelihoods for the rate ratio 122
14	Confounding and standardization 133
15	Comparison of rates within strata 141
16	Case-control studies 153
17	Likelihoods for the odds ratio 166
18	Comparison of odds within strata 175
19	Individually matched case-control studies 186
20	Tests for trend 197 *
21	The size of investigations 205 *

## II REGRESSION MODELS

22	Introduction to regression models	217
23	Poisson and logistic regression	227
24	Testing hypotheses	237
25	Models for dose-response	249
26	More about interaction	261
27	Choice and interpretation of models	271
★ 28	Additivity and synergism	282
29	Conditional logistic regression	290
30	Cox's regression analysis	298
31	Time-varying explanatory variables	307
★ 32	Three examples	319
33	Nested case-control studies	329
34	Gaussian regression models	336
35	Postscript	346

## III APPENDICES

A	Exponentials and logarithms	351
★ B	Some basic calculus	354
★ C	Approximate profile likelihoods	357
D	Table of the chi-squared distribution	363
	Index	365

★ Denotes a chapter which could be omitted from a first course.

## Part I

# Probability models and likelihood