

Statistics for Biology and Health

Series Editors

K. Dietz, M. Gail, K. Krickeberg, B. Singer

Springer Science+Business Media, LLC

Statistics for Biology and Health

Klein/Moeschberger: Survival Analysis: Techniques for Censored and Truncated Data.

Kleinbaum: Logistic Regression: A Self-Learning Text.

Kleinbaum: Survival Analysis: A Self-Learning Text.

Lange: Mathematical and Statistical Methods for Genetic Analysis.

Manton/Singer/Suzman: Forecasting the Health of Elderly Populations.

Salsburg: The Use of Restricted Significance Tests in Clinical Trials.

Kenneth Lange

Mathematical and Statistical Methods for Genetic Analysis



Springer

Kenneth Lange
Department of Biostatistics
and Mathematics
University of Michigan
Ann Arbor, MI 48109-2029
USA

Series Editors

K. Dietz
Institut für Medizinische Biometrie
Universität Tübingen
Westbahnhofstr. 55
D 72070 Tübingen
Germany

K. Krickeberg
3 Rue de L'Estrapade
75005 Paris
France

M. Gail
National Cancer Institute
Rockville, MD 20892
USA

B. Singer
Office of Population Research
Princeton University
Princeton, NJ 08544
USA

With 30 illustrations.

Library of Congress Cataloging-in-Publication Data
Lange, Kenneth.

Mathematical and statistical methods for genetic analysis /
Kenneth Lange.

p. cm. — (Statistics for biology and health)
Includes bibliographical references and index.

1. Genetics—Mathematics. 2. Genetics—Statistical methods.

I. Title. II. Series.

QH438.4.M33L36 1997

576.5'01'51—dc21

96-49533

© 1997 Springer Science+Business Media New York

Originally published by Springer-Verlag New York, Inc in 1997.

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer-Verlag Berlin Heidelberg GmbH), except for brief excerpts in connection with reviews or scholarly

analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use of general descriptive names, trade names, trademarks, etc., in this publication, even if the former are not especially identified, is not to be taken as a sign that such names, as understood by the Trade Marks and Merchandise Marks Act, may accordingly be used freely by anyone.

Production managed by Timothy Taylor; manufacturing supervised by Johanna Tschebull.

Camera-ready copy prepared from the author's LaTeX files.

Printed and bound by Maple-Vail Book Manufacturing Group, York, PA.

9 8 7 6 5 4 3 2 1

ISBN 978-1-4757-2741-8 ISBN 978-1-4757-2739-5 (eBook)

DOI 10.1007/978-1-4757-2739-5

Softcover reprint of the harcover 1st edition 1997

To Genie

Preface

When I was a postdoctoral fellow at UCLA more than two decades ago, I learned genetic modeling from the delightful texts of Elandt-Johnson [2] and Cavalli-Sforza and Bodmer [1]. In teaching my own genetics course over the past few years, first at UCLA and later at the University of Michigan, I longed for an updated version of these books. Neither appeared and I was left to my own devices. As my hastily assembled notes gradually acquired more polish, it occurred to me that they might fill a useful niche. Research in mathematical and statistical genetics has been proceeding at such a breathless pace that the best minds in the field would rather create new theories than take time to codify the old. It is also far more profitable to write another grant proposal. Needless to say, this state of affairs is not ideal for students, who are forced to learn by wading unguided into the confusing swamp of the current scientific literature.

Having set the stage for nobly rescuing a generation of students, let me inject a note of honesty. This book is not the monumental synthesis of population genetics and genetic epidemiology achieved by Cavalli-Sforza and Bodmer. It is also not the sustained integration of statistics and genetics achieved by Elandt-Johnson. It is not even a compendium of recommendations for carrying out a genetic study, useful as that may be. My goal is different and more modest. I simply wish to equip students already sophisticated in mathematics and statistics to engage in genetic modeling. These are the individuals capable of creating new models and methods for analyzing genetic data. No amount of expertise in genetics can overcome mathematical and statistical deficits. Conversely, no mathematician or statistician ignorant of the basic principles of genetics can ever hope to identify worthy problems. Collaborations between geneticists on one side and mathematicians and statisticians

on the other can work, but it takes patience and a willingness to learn a foreign vocabulary.

So what are my expectations of readers and students? This is a hard question to answer, in part because the level of the mathematics required builds as the book progresses. At a minimum, readers should be familiar with notions of theoretical statistics such as likelihood and Bayes' theorem. Calculus and linear algebra are used throughout. The last few chapters make fairly heavy demands on skills in theoretical probability and combinatorics. For a few subjects such as continuous time Markov chains and Poisson approximation, I sketch enough of the theory to make the exposition of applications self-contained. Exposure to interesting applications should whet students' appetites for self-study of the underlying mathematics. Everything considered, I recommend that instructors cover the chapters in the order indicated and determine the speed of the course by the mathematical sophistication of the students. There is more than ample material here for a full semester, so it is pointless to rush through basic theory if students encounter difficulty early on. Later chapters can be covered at the discretion of the instructor.

The matter of biological requirements is also problematic. Neither the brief review of population genetics in Chapter 1 nor the primer of molecular genetics in the Appendix is a substitute for a rigorous course in modern genetics. Although many of my classroom students have had little prior exposure to genetics, I have always insisted that those intending to do research fill in the gaps in their knowledge. Students in the mathematical sciences occasionally complain to me that learning genetics is hopeless because the field is in such rapid flux. While I am sympathetic to the difficult intellectual hurdles ahead of them, this attitude is a prescription for failure. Although genetics lacks the theoretical coherence of mathematics, there are fundamental principles and crucial facts that will never change. My advice is follow your curiosity and learn as much genetics as you can. In scientific research chance always favors the well prepared.

The incredible flowering of mathematical and statistical genetics over the past two decades makes it impossible to summarize the field in one book. I am acutely aware of my failings in this regard, and it pains me to exclude most of the history of the subject and to leave unmentioned so many important ideas. I apologize to my colleagues. My own work receives too much attention; my only excuse is that I understand it best. Fortunately, the recent book of Michael Waterman delves into many of the important topics in molecular genetics missing here [4].

I have many people to thank for helping me in this endeavor. Carol Newton nurtured my early career in mathematical biology and encouraged me to write a book in the first place. Daniel Weeks and Eric Sobel deserve special credit for their many helpful suggestions for improving the text. My genetics colleagues David Burke, Richard Gatti, and Miriam Meisler read and corrected my first draft of the appendix. David Cox, Richard Gatti, and James Lake kindly contributed data. Janet Sinsheimer and Hongyu Zhao provided numerical examples for Chapters 10 and 12, respectively. Many students at UCLA and Michigan checked the problems and proofread the text. Let me single out Ru-zong Fan, Ethan Lange, Laura Lazzeroni, Eric Schadt, Janet Sinsheimer, Heather Stringham, and Wynn Walker for their

diligence. David Hunter kindly prepared the index. Doubtless a few errors remain, and I would be grateful to readers for their corrections. Finally, I thank my wife Genie, to whom I dedicate this book, for her patience and love.

A Few Words about Software

This text contains several numerical examples that rely on software from the public domain. Readers interested in a copy of the programs MENDEL and FISHER mentioned in Chapters 7 and 8 and the optimization program SEARCH used in Chapter 3 should get in touch with me. Laura Lazzeroni distributes software for testing transmission association and linkage disequilibrium as discussed in Chapter 4. Daniel Weeks is responsible for the software implementing the APM method of linkage analysis featured in Chapter 6. He and Eric Sobel also distribute software for haplotyping and stochastic calculation of location scores as covered in Chapter 9. Readers should contact Eric Schadt or Janet Sinsheimer for the phylogeny software of Chapter 10 and Michael Boehnke for the radiation hybrid software of Chapter 11. Further free software for genetic analysis is listed in the recent book by Ott and Terwilliger [3].

References

- [1] Cavalli-Sforza LL, Bodmer WF (1971) *The Genetics of Human Populations*. Freeman, San Francisco
- [2] Elandt-Johnson RC (1971) *Probability Models and Statistical Methods in Genetics*. Wiley, New York
- [3] Terwilliger JD, Ott J (1994) *Handbook of Human Genetic Linkage*. Johns Hopkins University Press, Baltimore
- [4] Waterman MS (1995) *Introduction to Computational Biology: Maps, Sequences, and Genomes*. Chapman and Hall, London

Acknowledgments

- [1] Lange K, Boehnke M (1992) Bayesian methods and optimal experimental design for gene mapping by radiation hybrids. *Ann Hum Genet* 56:119–144
- [2] Goradia TM, Lange K, Miller PL, Nadkarni PM (1992) Fast computation of genetic likelihoods on human pedigree data. *Hum Hered* 42:42–62
- [3] Lange K, Sinsheimer JS (1992) Calculation of genetic identity coefficients. *Ann Hum Genet* 56:339–346
- [4] Sobel E, Lange K (1993) Metropolis sampling in pedigree analysis. *Stat Methods Med Res* 2:263–282
- [5] Lange K (1995) Applications of the Dirichlet distribution to forensic match probabilities. *Genetica* 96:107–117
- [6] Lange K, Boehnke M, Cox DR, Lunetta KL (1995) Statistical methods for polyploid radiation hybrid mapping. *Genome Res* 5:136–150
- [7] Sobel E, Lange K (1996) Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker sharing statistics. *Amer J Hum Genet* 58:1323–1337

Contents

Preface	vi
1 Basic Principles of Population Genetics	1
1.1 Introduction	1
1.2 Genetics Background	1
1.3 Hardy-Weinberg Equilibrium	4
1.4 Linkage Equilibrium	8
1.5 Selection	9
1.6 Balance Between Mutation and Selection	11
1.7 Problems	12
References	17
2 Counting Methods and the EM Algorithm	19
2.1 Introduction	19
2.2 Gene Counting	19
2.3 Description of the EM Algorithm	21
2.4 Ascent Property of the EM Algorithm	22
2.5 Allele Frequency Estimation by the EM Algorithm	24
2.6 Classical Segregation Analysis by the EM Algorithm	25
2.7 Problems	28
References	33
3 Newton's Method and Scoring	35
3.1 Introduction	35

3.2	Newton's Method	35
3.3	Scoring	36
3.4	Application to the Design of Linkage Experiments	39
3.5	Quasi-Newton Methods	42
3.6	The Dirichlet Distribution	43
3.7	Empirical Bayes Estimation of Allele Frequencies	44
3.8	Problems	47
	References	50
4	Hypothesis Testing and Categorical Data	52
4.1	Introduction	52
4.2	Hypotheses About Genotype Frequencies	52
4.3	Other Multinomial Problems in Genetics	55
4.4	The Z_{\max} Test	56
4.5	The W_d Statistic	58
4.6	Exact Tests of Independence	60
4.7	The Transmission/Disequilibrium Test	62
4.8	Problems	64
	References	67
5	Genetic Identity Coefficients	70
5.1	Introduction	70
5.2	Kinship and Inbreeding Coefficients	70
5.3	Condensed Identity Coefficients	74
5.4	Generalized Kinship Coefficients	75
5.5	From Kinship to Identity Coefficients	75
5.6	Calculation of Generalized Kinship Coefficients	77
5.7	Problems	80
	References	83
6	Applications of Identity Coefficients	85
6.1	Introduction	85
6.2	Genotype Prediction	86
6.3	Covariances for a Quantitative Trait	87
6.4	Risk Ratios and Genetic Model Discrimination	90
6.5	An Affecteds-Only Method of Linkage Analysis	93
6.6	Problems	97
	References	100
7	Computation of Mendelian Likelihoods	102
7.1	Introduction	102
7.2	Mendelian Models	102
7.3	Genotype Elimination and Allele Consolidation	105
7.4	Array Transformations and Iterated Sums	107
7.5	Array Factoring	109

7.6	Examples of Pedigree Analysis	111
7.7	Problems	117
	References	121
8	The Polygenic Model	123
8.1	Introduction	123
8.2	Maximum Likelihood Estimation by Scoring	124
8.3	Application to Gc Measured Genotype Data	128
8.4	Multivariate Traits	129
8.5	Left and Right-Hand Finger Ridge Counts	130
8.6	The Hypergeometric Polygenic Model	131
8.7	Application to Risk Prediction	135
8.8	Problems	136
	References	139
9	Markov Chain Monte Carlo Methods	142
9.1	Introduction	142
9.2	Review of Discrete-Time Markov Chains	143
9.3	The Hastings-Metropolis Algorithm and Simulated Annealing	146
9.4	Descent States and Descent Graphs	148
9.5	Descent Trees and the Founder Tree Graph	150
9.6	The Descent Graph Markov Chain	153
9.7	Computing Location Scores	156
9.8	Finding a Legal Descent Graph	157
9.9	Haplotyping	158
9.10	Application to Episodic Ataxia	158
9.11	Problems	159
	References	161
10	Reconstruction of Evolutionary Trees	164
10.1	Introduction	164
10.2	Evolutionary Trees	165
10.3	Maximum Parsimony	166
10.4	Review of Continuous-Time Markov Chains	169
10.5	A Nucleotide Substitution Model	171
10.6	Maximum Likelihood Reconstruction	174
10.7	Origin of the Eukaryotes	175
10.8	Problems	178
	References	181
11	Radiation Hybrid Mapping	183
11.1	Introduction	183
11.2	Models for Radiation Hybrids	184
11.3	Minimum Obligate Breaks Criterion	185
11.4	Maximum Likelihood Methods	187

11.5	Application to Haploid Data	190
11.6	Polyploid Radiation Hybrids	191
11.7	Maximum Likelihood Under Polyploidy	191
11.8	Obligate Breaks Under Polyploidy	195
11.9	Bayesian Methods	196
11.10	Application to Diploid Data	199
11.11	Problems	201
	References	204
12	Models of Recombination	206
12.1	Introduction	206
12.2	Mather's Formula and Its Generalization	207
12.3	Count-Location Model	209
12.4	Stationary Renewal Models	211
12.5	Poisson-Skip Model	213
12.6	Chiasma Interference	219
12.7	Application to <i>Drosophila</i> Data	221
12.8	Problems	223
	References	225
13	Poisson Approximation	228
13.1	Introduction	228
13.2	Poisson Approximation to the W_d Statistic	229
13.3	Construction of Somatic Cell Hybrid Panels	230
13.4	Biggest Marker Gap	232
13.5	Randomness of Restriction Sites	234
13.6	DNA Sequence Matching	237
13.7	Problems	240
	References	243
	Appendix: Molecular Genetics in Brief	245
A.1	Genes and Chromosomes	245
A.2	From Gene to Protein	246
A.3	Manipulating DNA	248
A.4	Mapping Strategies	250
	References	252
	Index	255