# Statistics for Biology and Health

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

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

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