NSF-CBMS Regional Conference Series
in Probability and Statistics
Volume 6

STATISTICAL INFERENCE FROM GENETIC DATA ON PEDIGREES

Elizabeth A. Thompson
University of Washington

Institute of Mathematical Statistics
Beachwood, Ohio

American Statistical Association
Alexandria, Virginia

Conference Board of the Mathematical Sciences

Regional Conference Series
in Probability and Statistics

Supported by the
National Science Foundation

The production of the *NSF-CBMS Regional Conference Series in Probability and Statistics* is managed by the Institute of Mathematical Statistics: Barry Arnold, IMS Managing Editor, Statistics; Patrick Kelly, IMS Production Editor; Julia Norton, IMS Treasurer; and Elyse Gustafson, IMS Executive Director.

Library of Congress Control Number: 00-134575

International Standard Book Number 0-940600-49-8

Printed in the United States of America

Contents

List of Tables

List of Figures

Preface

This monograph is based primarily on material presented at the CBMS Summer Course on **Inferences from genetic data on pedigrees** given at Michigan Technical University, Hougton, Michigan, in July 1999. This monograph is not a textbook; it contains no exercises, and is insufficiently detailed for that purpose. However, it could be used as a textbook, either in conjunction with the excellent texts of Weir (1996), Lange (1997) and Ott (1999), or by advanced students who will consult the cited literature for details.

The notes used at the Summer Course have been augmented by material from two lecture classes given at the University of Washington. A Special Topics class was given in January-March, 1999, and additional background on Markov chain Monte Carlo and Monte Carlo EM are included from that class. Some details were also first presented at a SEMSTAT workshop in Eindhoven in March 1999 (Thompson, 2000*b*). Although material has been added, the examples in Chapter 10 and on identity by descent under interference (section 11.2) were first presented at a Royal Statistical Society Meeting in London, in March 1999 (Thompson, 2000*a*). Versions of Figures 9.1, 10.1, 10.2, 10.6, and 10.7 first appeared in Thompson (2000*a*). However, the 11-chapter monograph follows closely the ten sessions of the Summer Course presentations, with chapter 2 being the only addition, providing statistical background with genetic examples. The order of Chapters 8 and 9 has been reversed from the Summer Course; a case can be made for either ordering.

A more basic Statistical Genetics class was given in Fall 1999, at University of Washington, and led to extensive revision of Chapters 1-4. It is hoped that the monograph can thus serve two purposes. For example, a more introductory course could cover of Chapters 1-4, with final material taken from sections 6.1, 6.2, 7.1, and 7.2. More advanced students could skip Chapters 1-2, skim Chapters 3-5, and study the later chapters more thoroughly.

I would like to thank Dr.Anant Godbole and Dr.JianPing Dong, for their excellent organization of the CBMS Regional Research Conference at Michigan Technical University. I am also grateful to the many students who attended this course, and to students attending the two University of Washington courses, for their helpful comments and criticisms. In particular, I would like to thank Eric Anderson, Nicky Chapman and Dr.Ellen Wijsman for help with LaTeX, BibTeX, Xfig, and GENEHUNTER, and for many discussions. I am grateful to Amy Anderson for her thorough and critical reading of Chapters 1 to 5, and to Eric Anderson, Dr.Erin Conlon, Dr.Mary Kuhner, Anne-Louise Leutenegger, and Jessica

Maia, who all read and commented on other chapters.

Some of the MCMC work was undertaken in collaboration with Dr.Simon Heath. In particular, the implementation of the algorithm described in section 3.6 and the initial incorporation of the L-sampler of Heath (1997) into our M-sampler software to create the LM-sampler (section 10.6) are both due to Dr.Heath. Figures 1.1, 1.2, 3.4, 3.5, and 10.3, first appeared in Thompson and Heath (1999), and are also due to Dr.Heath. I am grateful to Dr.Heath for our continuing collaboration.

The CBMS Regional Research Conference was funded by NSF grant number 98-13767 to Dr.Jianping Dong and Dr.Anant Godbole of The Mathematical Sciences Department of Michigan Technical University, Houghton, MI.

Table of Notation

Since there are an insufficient number of user-friendly letters and symbols, some must be used for more than one purpose. However, for convenience, we summarize the principal usages here

Notation	Usage
Parameters	
θ	the general (set of) parameters of a model
ρ	a recombination frequency parameter
γ	a (trait) locus location
Γ_M	a marker map; set of arker locations
β	a trait model penetrance parameter
r	number of multinomial outcomes (or phenotypes)
$p_1, \ldots, p_r$	probabilities of multinomial outcomes
k	number of alleles at a locus
$q_1, \ldots, q_k$	population allele frequencies at this locus
q	an allele frequency, often for a recessive allele
ψ	a kinship coefficient
ϕ	chiasmata avoidance function
$\kappa_i, i = 0, 1, 2$	gene-identity probabilities
Indices and labels	
i	an index used primarily for individuals or meioses
j	an index used primarily for alleles or loci
k, k_i	a label for a gene
L	a number of loci ordered on a chromosome
m	a count, often of the number of meioses
v	miscellaneous other counts, of genes for example
n	sample size
F	father, or paternal, often as subscript
M	mother, or maternal, often as subscript
	also marker, as in marker data $\mathbf{Y}_M$
N	Monte Carlo sample size
	also (Chapter 5) the random number of chiasmata)
τ	an index of Monte Carlo or MCMC realizations
$\mathcal{T}$	a set of indices of latent variables
$\mathcal{D}$	a set of indices of data observations
Variables	
$A_1, ..., A_k$	the alleles at a locus
$\mathbf{Y}$, value $\mathbf{y}$	the data random variables (usually phenotypes)
$\mathbf{Y}_M$	phenotypes at marker loci, in linkage mapping
$\mathbf{Y}_T$	trait phenotypes; $\mathbf{Y} = (\mathbf{Y}_T, \mathbf{Y}_M)$
$\mathbf{X}$, value $\mathbf{x}$	latent variables
$\mathbf{X}^\dagger$	a proposed value of $\mathbf{X}$ in Monte Carlo sampling
$\mathbf{X}^*$	a sampled or resampled value of $\mathbf{X}$ in Monte Carlo
$\mathbf{G} = \{G_i\}$	the set of genotypes of individuals i
g	a genotype — a possible value of G_i

Notation	Usage
Variables continued	
$\mathbf{S} = \{S_{i,j}\}$	set of meiosis indicators for meioses i and loci j
$S_{\bullet,j}$	the vector of $S_{i,j}$ at given locus j
$S_{i,\bullet}$	the vector of $S_{i,j}$ at given meiosis i
$G_{\bullet,j}, G_{i,\bullet}$	similarly for genotypes, locus j, individual i
$Y_{\bullet,j}, Y_{i,\bullet}$	similarly for phenotypes, locus j, individual i
$Y^{(j)}$	the data $\{Y_{\bullet,1}, Y_{\bullet,2}, \ldots, Y_{\bullet,j}\}$; $\mathbf{Y} = Y^{(L)}$
$\mathbf{J} = \mathbf{J}(\mathbf{S})$	a gene *ibd* pattern, a function of $\mathbf{S}$
$I_1, ..., I_{L-1}$	the intervals between L ordered loci
$\mathbf{R} = (R_j; j = 1, ..., L-1)$	the recombination indicators in intervals I_j
$\mathbf{r}$	a vector of recombination indices; value of $\mathbf{R}$
$\mathbf{C} = (C_j; j = 1, ..., L-1)$	the chiasmata presence/absence indicators in intervals I_j
$\mathbf{c}$	a vector of chiasma indices; value of $\mathbf{C}$
T, value t	a count (often binomial)
t_j	a multinomial count, e.g. of latent genotypes; also (Chapter 5) a set of binary indicators
n_{jl}, n_j	multinomial data counts, of observable phenotypes or genotypes
m_j	multinomial counts, often of alleles
Functions and probabilities	
Pr	probability, when not indexed by a parameter
$\Pr(E; \theta)$	probability of event E under model θ
$P_\theta(\cdot)$	a probability distribution, indexed by θ
$P^*(\cdot)$	a probability distribution, used for the sampling or resampling distribution in Monte Carlo methods
$\mathrm{E}_\theta(\cdot)$	Expectation, under a model indexed by θ
$\Phi(\cdot)$	the standard Normal (Gaussian) cumulative distribution function
$I(\cdot)$	the indicator function of an event
$L(\theta)$ or $L_{\mathbf{y}}(\theta)$	the likelihood for parameter θ given data $\mathbf{y}$
$L(\theta; \mathbf{Y})$	the likelihood function, considered also as a function of data random variables $\mathbf{Y}$
ℓ or $\ell(\theta)$	the log-likelihood function for parameter θ
$K_n(\theta; \theta_0)$	Kullback-Leibler information in a sample size n
$K_{\mathbf{y}}(\theta; \theta_0)$	K-L information in latent $\mathbf{X}$ given data $\mathbf{y}$
$H_{\mathbf{y}}(\theta; \theta_0)$	expected complete-data log-likelihood given $\mathbf{Y} = \mathbf{y}$: $\mathrm{E}_{\theta_0}(\log P_\theta(\mathbf{X}, \mathbf{Y}) \mid \mathbf{Y} = \mathbf{y})$
$R(\cdot)$ and $R^*(\cdot)$	cumulative probabilities of data used in computing probabilities on graphs or pedigrees
$Q(\cdot)$, $Q^*(\cdot)$, $Q^\dagger(\cdot)$	cumulative conditional probabilities of latent variables given data on graphs or pedigrees
$h(\mathbf{X}^\dagger; \mathbf{X})$	Hastings ratio for proposed $\mathbf{X}^\dagger$ when at state $\mathbf{X}$
$q(\mathbf{X}^\dagger; \mathbf{X})$	proposal probability for $\mathbf{X}^\dagger$ when at state $\mathbf{X}$
a	the Metropolis-Hastings acceptance probability