Lecture Notes in Mathematics

Simon Tavaré
Ofer Zeitouni

Lectures on Probability Theory and Statistics

1837

Ecole d'Eté de Probabilités de Saint-Flour XXXI – 2001

Editor: J. Picard

Springer

Simon Tavaré Ofer Zeitouni

Lectures on Probability Theory and Statistics

Ecole d'Eté de Probabilités de Saint-Flour XXXI - 2001

Editor: Jean Picard

Springer

Authors

Simon Tavaré
Program in Molecular and
Computational Biology
Department of Biological Sciences
University of Southern California
Los Angeles, CA 90089-1340
USA

e-mail: stavare@usc.edu

Ofer Zeitouni
Departments of Electrical Engineering
and of Mathematics
Technion - Israel Institute of Technology
Haifa 32000, Israel
and
Department of Mathematics
University of Minnesota
206 Church St. SE
Minneapolis, MN 55455
USA

e-mail: zeitouni@ee.technion.ac.il
zeitouni@math.umn.edu

Editor

Jean Picard
Laboratoire de Mathématiques Appliquées
UMR CNRS 6620
Université Blaise Pascal Clermont-Ferrand
63177 Aubière Cedex, France

e-mail: Jean.Picard@math.univ-bpclermont.fr

Cover illustration: Blaise Pascal (1623-1662)

Cataloging-in-Publication Data applied for
Bibliographic information published by Die Deutsche Bibliothek

Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the Internet at http://dnb.ddb.de

Mathematics Subject Classification (2001):
60-01, 60-06, 62-01, 62-06, 92D10, 60K37, 60F05, 60F10

ISSN 0075-8434 Lecture Notes in Mathematics
ISSN 0721-5363 Ecole d'Eté des Probabilits de St. Flour
ISBN 3-540-20832-1 Springer-Verlag Berlin Heidelberg New York

Springer-Verlag Berlin Heidelberg New York a member of BertelsmannSpringer
Science + Business Media GmbH

http://www.springer.de

Printed in Germany

Typesetting: Camera-ready TEX output by the authors

SPIN: 10981573 41/3142/du - 543210 - Printed on acid-free paper

Preface

Three series of lectures were given at the 31st Probability Summer School in Saint-Flour (July 8–25, 2001), by the Professors Catoni, Tavaré and Zeitouni. In order to keep the size of the volume not too large, we have decided to split the publication of these courses into two parts. This volume contains the courses of Professors Tavaré and Zeitouni. The course of Professor Catoni entitled "Statistical Learning Theory and Stochastic Optimization" will be published in the *Lecture Notes in Statistics*. We thank all the authors warmly for their important contribution.

55 participants have attended this school. 22 of them have given a short lecture. The lists of participants and of short lectures are enclosed at the end of the volume.

Finally, we give the numbers of volumes of Springer *Lecture Notes* where previous schools were published.

Lecture Notes in Mathematics

1971: vol 307	1973: vol 390	1974: vol 480	1975: vol 539
1976: vol 598	1977: vol 678	1978: vol 774	1979: vol 876
1980: vol 929	1981: vol 976	1982: vol 1097	1983: vol 1117
1984: vol 1180	1985/86/87: vol 1362	1988: vol 1427	1989: vol 1464
1990: vol 1527	1991: vol 1541	1992: vol 1581	1993: vol 1608
1994: vol 1648	1995: vol 1690	1996: vol 1665	1997: vol 1717
1998: vol 1738	1999: vol 1781	2000: vol 1816	

Lecture Notes in Statistics

1986: vol 50	2003: vol 179

Contents

Part I Simon Tavaré: Ancestral Inference in Population Genetics

Part II Ofer Zeitouni: Random Walks in Random Environment

Part I

Simon Tavaré: Ancestral Inference in Population Genetics

Ancestral Inference in Population Genetics

Simon Tavaré

Departments of Biological Sciences, Mathematics and Preventive Medicine
University of Southern California.

1 Introduction

One of the most important challenges facing modern biology is how to make sense of genetic variation. Understanding how genotypic variation translates into phenotypic variation, and how it is structured in populations, is fundamental to our understanding of evolution. Understanding the genetic basis of variation in phenotypes such as disease susceptibility is of great importance to human geneticists. Technological advances in molecular biology are making it possible to survey variation in natural populations on an enormous scale. The most dramatic examples to date are provided by Perlegen Sciences Inc., who resequenced 20 copies of chromosome 21 (Patil *et al.*, 2001) and by Genaissance Pharmaceuticals Inc., who studied haplotype variation and linkage disequilibrium across 313 human genes (Stephens *et al.*, 2001). These are but two of the large number of variation surveys now underway in a number of organisms. The amount of data these studies will generate is staggering, and the development of methods for their analysis and interpretation has become central. In these notes I describe the basics of *coalescent theory*, a useful quantitative tool in this endeavor.

1.1 Genealogical processes

These Saint Flour lectures concern *genealogical processes*, the stochastic models that describe the ancestral relationships among samples of individuals. These individuals might be species, humans or cells – similar methods serve to analyze and understand data on very disparate time scales. The main theme is an account of methods of statistical inference for such processes, based primarily on stochastic computation methods. The notes do not claim to be even-handed or comprehensive; rather, they provide a personal view of some of the theoretical and computational methods that have arisen over the last 20 years. A comprehensive treatment is impossible in a field that is evolving as fast as this one. Nonetheless I think the notes serve as a useful starting point for accessing the extensive literature.

Understanding molecular variation data

The first lecture in the Saint Flour Summer School series reviewed some basic molecular biology and outlined some of the problems faced by computational molecular biologists. This served to place the problems discussed in the remaining lectures into a broader perspective. I have found the books of Hartl and Jones (2001) and Brown (1999) particularly useful.

It is convenient to classify evolutionary problems according to the time scale involved. On long time scales, think about trying to reconstruct the molecular phylogeny of a collection of species using DNA sequence data taken

from a homologous region in each species. Not only is the phylogeny, or branching order, of the species of interest but so too might be estimation of the divergence time between pairs of species, of aspects of the mutation process that gave rise to the observed differences in the sequences, and questions about the nature of the common ancestor of the species. A typical population genetics problem involves the use of patterns of variation observed in a sample of humans to locate disease susceptibility genes. In this example, the time scale is of the order of thousands of years. Another example comes from cancer genetics. In trying to understand the evolution of tumors we might extract a sample of cells, type them for microsatellite variation at a number of loci and then use the observed variability to infer the time since a checkpoint in the tumor's history. The time scale in this example is measured in years.

The common feature that links these examples is the dependence in the data generated by common ancestral history. Understanding the way in which ancestry produces dependence in the sample is the key principle of these notes. Typically the ancestry is never known over the whole time scale involved. To make any progress, the ancestry has to be modelled as a stochastic process. Such processes are the subject of these notes.

Backwards or Forwards?

The theory of population genetics developed in the early years of the last century focused on a *prospective* treatment of genetic variation (see Provine (2001) for example). Given a stochastic or deterministic model for the evolution of gene frequencies that allows for the effects of mutation, random drift, selection, recombination, population subdivision and so on, one can ask questions like 'How long does a new mutant survive in the population?', or 'What is the chance that an allele becomes fixed in the population?'. These questions involve the analysis of the future behavior of a system given initial data. Most of this theory is much easier to think about if the focus is *retrospective*. Rather than ask where the population will go, ask where it has been. This changes the focus to the study of ancestral processes of various sorts. While it might be a truism that genetics is all about ancestral history, this fact has not pervaded the population genetics literature until relatively recently. We shall see that this approach makes most of the underlying methodology easier to derive – essentially all classical prospective results can be derived more simply by this dual approach – and in addition provides methods for analyzing modern genetic data.

1.2 Organization of the notes

The notes begin with forwards and backwards descriptions of the Wright-Fisher model of gene frequency fluctuation in Section 2. The ancestral process that records the number of distinct ancestors of a sample back in time is described, and a number of its basic properties derived. Section 3 introduces

the effects of mutation in the history of a sample, introduces the genealogical approach to simulating samples of genes. The main result is a derivation of the Ewens sampling formula and a discussion of its statistical implications. Section 4 introduces Kingman's coalescent process, and discusses the robustness of this process for different models of reproduction.

Methods more suited to the analysis of DNA sequence data begin in Section 5 with a theoretical discussion of the infinitely-many-sites mutation model. Methods for finding probabilities of the underlying reduced genealogical trees are given. Section 6 describes a computational approach based on importance sampling that can be used for maximum likelihood estimation of population parameters such as mutation rates. Section 7 introduces a number of problems concerning inference about properties of coalescent trees conditional on observed data. The motivating example concerns inference about the time to the most recent common ancestor of a sample. Section 8 develops some theoretical and computational methods for studying the ages of mutations. Section 9 discusses Markov chain Monte Carlo approaches for Bayesian inference based on sequence data. Section 10 introduces Hudson's coalescent process that models the effects of recombination. This section includes a discussion of ancestral recombination graphs and their use in understanding linkage disequilibrium and haplotype sharing.

Section 11 discusses some alternative approaches to inference using approximate Bayesian computation. The examples include two at opposite ends of the evolutionary time scale: inference about the divergence time of primates and inference about the age of a tumor. This section includes a brief introduction to computational methods of inference for samples from a branching process. Section 12 concludes the notes with pointers to some topics discussed in the Saint Flour lectures, but not included in the printed version. This includes models with selection, and the connection between the stochastic structure of certain decomposable combinatorial models and the Ewens sampling formula.

1.3 Acknowledgements

Paul Marjoram, John Molitor, Duncan Thomas, Vincent Plagnol, Darryl Shibata and Oliver Will were involved with aspects of the unpublished research described in Section 11. I thank Lada Markovtsova for permission to use some of the figures from her thesis (Markovtsova (2000)) in Section 9. I thank Magnus Nordborg for numerous discussions about the mysteries of recombination. Above all I thank Warren Ewens and Bob Griffiths, collaborators for over 20 years. Their influence on the statistical development of population genetics has been immense; it is clearly visible in these notes.

Finally I thank Jean Picard for the invitation to speak at the summer school, and the Saint-Flour participants for their comments on the earlier version of the notes.

2 The Wright-Fisher model

This section introduces the Wright-Fisher model for the evolution of gene frequencies in a finite population. It begins with a prospective treatment of a population in which each individual is one of two types, and the effects of mutation, selection, ... are ignored. A genealogical (or retrospective) description follows. A number of properties of the ancestral relationships among a sample of individuals are given, along with a genealogical description in the case of variable population size.

2.1 Random drift

The simplest Wright-Fisher model (Fisher (1922), Wright (1931)) describes the evolution of a two-allele locus in a population of constant size undergoing random mating, ignoring the effects of mutation or selection. This is the so-called 'random drift' model of population genetics, in which the fundamental source of "randomness" is the reproductive mechanism.

A Markov chain model

We assume that the population is of constant size N in each non-overlapping generation n, $n = 0, 1, 2, \ldots$. At the locus in question there are two alleles, denoted by A and B. X_n counts the number of A alleles in generation n. We assume first that there is no mutation between the types. The population at generation $r+1$ is derived from the population at time r by binomial sampling of N genes from a gene pool in which the fraction of A alleles is its current frequency, namely $\pi_i = i/N$. Hence given $X_r = i$, the probability that $X_{r+1} = j$ is

$$p_{ij} = \binom{N}{j} \pi_i^j (1 - \pi_i)^{N-j}, \quad 0 \le i, j \le N. \tag{2.1.1}$$

The process $\{X_r, r = 0, 1, \ldots\}$ is a time-homogeneous Markov chain. It has transition matrix $P = (p_{ij})$, and state space $\mathcal{S} = \{0, 1, \ldots, N\}$. The states 0 and N are absorbing; if the population contains only one allele in some generation, then it remains so in every subsequent generation. In this case, we say that the population is *fixed* for that allele.

The binomial nature of the transition matrix makes some properties of the process easy to calculate. For example,

$$\mathbb{E}(X_r | X_{r-1}) = N \frac{X_{r-1}}{N} = X_{r-1},$$

so that by averaging over the distribution of X_{r-1} we get $\mathbb{E}(X_r) = \mathbb{E}(X_{r-1})$, and

$$\mathbb{E}(X_r) = \mathbb{E}(X_0), \quad r = 1, 2, \ldots. \tag{2.1.2}$$

The result in (2.1.2) can be thought of as the analog of the Hardy-Weinberg law: in an infinitely large random mating population, the relative frequency of the alleles remains constant in every generation. Be warned though that average values in a stochastic process do not tell the whole story! While on average the number of A alleles remains constant, variability must eventually be lost. That is, eventually the population contains all A alleles or all B alleles.

We can calculate the probability a_i that eventually the population contains only A alleles, given that $X_0 = i$. The standard way to find such a probability is to derive a system of equations satisfied by the a_i. To do this, we condition on the value of X_1. Clearly, $a_0 = 0, a_N = 1$, and for $1 \leq i \leq N-1$, we have

$$a_i = p_{i0} \cdot 0 + p_{iN} \cdot 1 + \sum_{j=1}^{N-1} p_{ij} a_j. \tag{2.1.3}$$

This equation is derived by noting that if $X_1 = j \in \{1, 2, \ldots, N-1\}$, then the probability of reaching N before 0 is a_j. The equation in (2.1.3) can be solved by recalling that $\mathbb{E}(X_1 \mid X_0 = i) = i$, or

$$\sum_{j=0}^{N} p_{ij} j = i.$$

It follows that $a_i = Ci$ for some constant C. Since $a_N = 1$, we have $C = 1/N$, and so $a_i = i/N$. Thus the probability that an allele will fix in the population is just its initial frequency.

The variance of X_r can also be calculated from the fact that

$$\mathrm{Var}(X_r) = \mathbb{E}(\mathrm{Var}(X_r | X_{r-1})) + \mathrm{Var}(\mathbb{E}(X_r | X_{r-1})).$$

After some algebra, this leads to

$$\mathrm{Var}(X_r) = \mathbb{E}(X_0)(N - \mathbb{E}(X_0))(1 - \lambda^r) + \lambda^r \mathrm{Var}(X_0), \tag{2.1.4}$$

where

$$\lambda = 1 - 1/N.$$

We have noted that genetic variability in the population is eventually lost. It is of some interest to assess how fast this loss occurs. A simple calculation shows that

$$\mathbb{E}(X_r(N - X_r)) = \lambda^r \mathbb{E}(X_0(N - X_0)). \tag{2.1.5}$$

Multiplying both sides by $2N^{-2}$ shows that the probability $h(r)$ that two genes chosen at random with replacement in generation r are different is

$$h(r) = \lambda^r h(0). \tag{2.1.6}$$

The quantity $h(r)$ is called the *heterozygosity* of the population in generation r, and it measures the genetic variability surviving in the population. Equation

(2.1.6) shows that the heterozygosity decays geometrically quickly as $r \to \infty$. Since fixation must occur, we have $h(r) \to 0$.

We have seen that variability is lost from the population. How long does this take? First we find an equation satisfied by m_i, the mean time to fixation starting from $X_0 = i$. To do this, notice first that $m_0 = m_N = 0$, and, by conditioning on the first step once more, we see that for $1 \leq i \leq N-1$

$$\begin{aligned} m_i &= p_{i0} \cdot 1 + p_{iN} \cdot 1 + \sum_{j=1}^{N-1} p_{ij}(1 + m_j) \\ &= 1 + \sum_{j=0}^{N} p_{ij} m_j. \end{aligned} \tag{2.1.7}$$

Finding an explicit expression for m_i is difficult, and we resort instead to an approximation when N is large and time is measured in units of N generations.

Diffusion approximations

This takes us into the world of diffusion theory. It is usual to consider not the total number $X_r \equiv X(r)$ of A alleles but rather the proportion X_r/N. To get a non-degenerate limit we must also rescale time, in units of N generations. This leads us to study the rescaled process

$$Y_N(t) = N^{-1} X(\lfloor Nt \rfloor), \quad t \geq 0, \tag{2.1.8}$$

where $\lfloor x \rfloor$ is the integer part of x. The idea is that as $N \to \infty$, $Y_N(\cdot)$ should converge in distribution to a process $Y(\cdot)$. The fraction $Y(t)$ of A alleles at time t evolves like a continuous-time, continuous state-space process in the interval $S = [0, 1]$. $Y(\cdot)$ is an example of a diffusion process. Time scalings in units proportional to N generations are typical for population genetics models appearing in these notes.

Diffusion theory is the basic tool of classical population genetics, and there are several good references. Crow and Kimura (1970) has a lot of the 'old style' references to the theory. Ewens (1979) and Kingman (1980) introduce the sampling theory ideas. Diffusions are also discussed by Karlin and Taylor (1980) and Ethier and Kurtz (1986), the latter in the measure-valued setting. A useful modern reference is Neuhauser (2001).

The properties of a one-dimensional diffusion $Y(\cdot)$ are essentially determined by the infinitesimal mean and variance, defined in the time-homogeneous case by

$$\begin{aligned} \mu(y) &= \lim_{h \to 0} h^{-1} \mathbb{E}(Y(t+h) - Y(t) \mid Y(t) = y), \\ \sigma^2(y) &= \lim_{h \to 0} h^{-1} \mathbb{E}((Y(t+h) - Y(t))^2 \mid Y(t) = y). \end{aligned}$$

For the discrete Wright-Fisher model, we know that given $X_r = i$, X_{r+1} is binomially distributed with number of trials N and success probability i/N. Hence

$$\mathbb{E}(X(r+1)/N - X(r)/N \mid X(r)/N = i/N) = 0,$$
$$\mathbb{E}((X(r+1)/N - X(r)/N)^2 \mid X(r)/N = i/N) = \frac{1}{N}\frac{i}{N}\left(1 - \frac{i}{N}\right),$$

so that for the process $Y(\cdot)$ that gives the proportion of allele A in the population at time t, we have

$$\mu(y) = 0, \qquad \sigma^2(y) = y(1-y), \ 0 < y < 1. \tag{2.1.9}$$

Classical diffusion theory shows that the mean time $m(x)$ to fixation, starting from an initial fraction $x \in (0,1)$ of the A allele, satisfies the differential equation

$$\frac{1}{2}x(1-x)m''(x) = -1, \quad m(0) = m(1) = 0. \tag{2.1.10}$$

This equation, the analog of (2.1.7), can be solved using partial fractions, and we find that

$$m(x) = -2(x\log x + (1-x)\log(1-x)), \quad 0 < x < 1. \tag{2.1.11}$$

In terms of the underlying discrete model, the approximation for the expected number m_i of generations to fixation, starting from i A alleles, is $m_i \approx Nm(i/N)$. If $i/N = 1/2$,

$$Nm(1/2) = (-2\log 2)N \approx 1.39N \text{ generations},$$

whereas if the A allele is introduced at frequency $1/N$,

$$Nm(1/N) = 2\log N \text{ generations}.$$

2.2 The genealogy of the Wright-Fisher model

In this section we consider the Wright-Fisher model from a genealogical perspective. In the absence of recombination, the DNA sequence representing the gene of interest is a copy of a sequence in the previous generation, that sequence is itself a copy of a sequence in the generation before that and so on. Thus we can think of the DNA sequence as an 'individual' that has a 'parent' (namely the sequence from which is was copied), and a number of 'offspring' (namely the sequences that originate as a copy of it in the next generation).

To study this process either forwards or backwards in time, it is convenient to label the individuals in a given generation as $1, 2, \ldots, N$, and let ν_i denote the number of offspring born to individual i, $1 \le i \le N$. We suppose that individuals have independent Poisson-distributed numbers of offspring,