Probability and Its Applications

Published in association with the Applied Probability Trust

Probability and Its Applications

Anderson: Continuous-Time Markov Chains.
Azencott/Dacunha-Castelle: Series of Irregular Observations.
Bass: Diffusions and Elliptic Operators.
Bass: Probabilistic Techniques in Analysis.
Berglund/Gentz: Noise-Induced Phenomena in Slow-Fast Dynamical Systems.
Biagini/Hu/Øksendal/Zhang: Stochastic Calculus for Fractional Brownian Motion and Applications.
Chen: Eigenvalues, Inequalities and Ergodic Theory.
Choi: ARMA Model Identification.
Costa/Fragoso/Marques: Discrete-Time Markov Jump Linear Systems.
Daley/Vere-Jones: An Introduction to the Theory of Point Processes Volume I: Elementary Theory and Methods, Second Edition.
de la Peña/Giné: Decoupling: From Dependence to Independence.
Durrett: Probability Models for DNA Sequence Evolution, Second Edition.
Galambos/Simonelli: Bonferroni-type Inequalities with Applications.
Gani (ed): The Craft of Probabilistic Modelling.
Grandell: Aspects of Risk Theory.
Gut: Stopped Random Walks.
Guyon: Random Fields on a Network.
Kallenberg: Foundations of Modern Probability, Second Edition.
Last/Brandt: Marked Point Processes on the Real Line.
Leadbetter/Lindgren/Rootzén: Extremes and Related Properties of Random Sequences and Processes.
Molchanov: Theory of Random Sets.
Nualart: The Malliavin Calculus and Related Topics.
Rachev/Rüschendorf: Mass Transportation Problems. Volume I: Theory.
Rachev/Rüschendorf: Mass Transportation Problems. Volume II: Applications.
Resnick: Extreme Values, Regular Variation and Point Processes.
Schmidli: Stochastic Control in Insurance.
Shedler: Regeneration and Networks of Queues.
Silvestrov: Limit Theorems for Randomly Stopped Stochastic Processes.
Thorisson: Coupling, Stationarity and Regeneration.
Todorovic: An Introduction to Stochastic Processes and Their Applications.

Richard Durrett

Probability Models for DNA Sequence Evolution

Second Edition

Springer

Richard Durrett
Department of Mathematics
Cornell University
Ithaca, NY 14853-4201
USA

ISBN: 978-0-387-78168-6 e-ISBN: 978-0-387-78169-3
DOI: 10.1007/978-0-387-78168-6

Library of Congress Control Number: 2008929756

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Preface

> "Mathematics seems to endow one with something like a new sense."
>
> Charles Darwin

The goal of population genetics is to understand how genetic variability is shaped by natural selection, demographic factors, and random genetic drift. The stochastic evolution of a DNA segment that experiences recombination is a complex process, so many analyses are based on simulations or use heuristic methods. However, when formulas are available, they are preferable because, when simple, they show the dependence of observed quantities on the underlying parameters and, even when complicated, can be used to compute exact answers in a much shorter time than simulations can give good approximations.

The goal of this book is to give an account of useful analytical results in population genetics, together with their proofs. The latter are omitted in many treatments, but are included here because the derivation often gives insight into the underlying mechanisms and may help others to find new formulas. Throughout the book, the theoretical results are developed in close connection with examples from the biology literature that illustrate the use of these results. Along the way, there are many numerical examples and graphs to illustrate the main conclusions. To help the reader navigate the book, we have divided the sections into a large number of subsections listed in the index, and further subdivided the text with bold-faced headings (as in this Preface).

This book is written for mathematicians and for biologists alike. With mathematicians in mind, we assume no knowledge of concepts from biology. Section 1.1 gives a rapid introduction to the basic terminology. Other explanations are given as concepts arise. For biologists, we explain mathematical notation and terminology as it arises, so the only *formal* prerequisite for biologists reading this book is a one-semester undergraduate course in probability and some familiarity with Markov chains and Poisson processes will be very useful. We have emphasized the word *formal* here, because to read and understand all of the proofs will require more than these simple prerequisites. On the other hand, the book has been structured so that proofs can be omitted.

What is in this book?

Chapter 1 begins with the theory of neutral evolution in a homogeneously mixing population of constant size. We introduce and study the discrete-time Wright-Fisher model, the continuous-time Moran model, the coalescent, which describes the genealogy of a nonrecombining segment of DNA, and two simplified models of mutation: the infinite alleles and infinite sites models. Based on these results, Chapter 2 introduces the problem of testing to see if observed DNA sequences are consistent with the assumptions of the "null model" underlying the theory developed in Chapter 1.

Chapters 3 through 6 confront the complications that come from relaxing the assumptions of the models in Chapter 1. This material, which filled two chapters in the first edition, has doubled in size and contains many results from the last five years. Chapter 3 introduces the ancestral recombination graph and studies the effect of recombination on genetic variability and the problem of estimating the rate at which recombination occurs. Chapter 4 investigates the influence of large family sizes, population size changes, and population subdivision in the form of island models on the genealogy of a sample. Chapter 5 concerns the more subtle behavior of the stepping stone model, which depicts a population spread across a geographical range, not grouped into distinct subpopulations. Finally, Chapter 6 considers various forms of natural selection: directional selection and hitchhiking, background selection and Muller's ratchet, and balancing selection.

Chapters 7 and 8, which are new in this edition, treat the previous topics from the viewpoint of diffusion processes, continuous stochastic processes that arise from letting the population size $N \to \infty$ and at the same time running time at rate $O(N)$. A number of analytical complications are associated with this approach, but, at least in the case of the one-dimensional processes considered in Chapter 7, the theory provides powerful tools for computing fixation probabilities, expected fixation time, and the site frequency spectrum. In contrast, the theory of multidimensional diffusions described in Chapter 8 is more of an art than a science. However, it offers significant insights into recombination, Hill-Robertson interference, and gene duplication.

Chapter 9 tackles the relatively newer, and less well-developed, study of the evolution of whole genomes by chromosomal inversions, reciprocal translocations, and genome duplication. This chapter is the least changed from the previous edition but has new results about when the parsimony method is effective, Bayesian estimation of genetic distance, and the midpoint problem.

In addition to the three topics just mentioned, there are a number of results covered here that do not appear in most other treatments of the subject (given here with the sections in which they appear): Fu's covariance matrix for the site frequency spectrum (2.1), the sequentially Markovian coalescent (3.4), the beta coalescent for large family sizes (4.1), Malécot's recursion for identity by descent and its study by Fourier analysis (5.2), the "continuous" (or long-range) stepping stone model (5.5–5.6), Muller's ratchet and Kondrashov's result for truncation selection (6.4), approximations for the effect of hitchhiking and recurrent selective sweeps (6.5–6.7), the Poisson random field model (7.11), fluctuating selection (7.12), a new approximate formula for the effect of Hill-Robertson interference (8.3), and a new result showing that the subfunctionalization explanation of gene duplication is extremely unlikely in large populations (8.6).

Having bragged about what I do cover, I must admit that this book has little to say about computationally intensive procedures. Some of these methods are mentioned along the way, and in some cases (e.g., Hudson's composite likelihood method for estimating recombination rates, and the Kim and Stephan test) we give some of the underlying mathematics. However, not being a user of these methods, I could not explain to you how to use them any better than I could tell you how to make a chocolate soufflé. As in the case of cooking, if you want to learn, you can find recipes on the Internet. A good place to start is `www.molpopgen.org`.

Mens rea

In response to criticisms of the first edition and the opinions of a half-dozen experts hired to read parts of the first draft of the second edition, I have worked hard to track down errors and clarify the discussion. Undoubtedly, there are bugs that remain to be fixed, five years from now in the third edition. Comments and complaints can be emailed to rtd1@cornell.edu. My web page `www.math.cornell.edu/~durrett` can be consulted for corrections.

Interdisciplinary work, of the type described in the book, is not easy and is often frustrating. Mathematicians think that it is trivial because, in many cases, the analysis does not involve developing new mathematics. Biologists find the "trivial" calculations confusing, that the simple models omit important details, and are disappointed by the insights they provide. Nonetheless, I think that important insights can be obtained when problems are solved analytically, rather than being conquered by complicated programs running for days on computer clusters.

I would like to thank the postdocs and graduate students who in recent years have joined me on the journey to the purgatory at the interface between probability and biology (in my case, genetics and ecology): Janet Best, Ben Chan, Arkendra De, Emilia Huerta-Sanchez, Yannet Interian, Nicholas Lanchier, Vlada Limic, Lea Popovic, Daniel Remenik, Deena Schmidt, and Jason Schweinsberg. I appreciate the patience of my current co-authors on this list as I ignored our joint projects, so that I could devote all of my energy to finishing this book.

As I write this, a January (2008) thaw is melting the snow in upstate New York, just in time so that my wife (and BFF) Susan can drive my younger son, Greg, back to MIT to start his fourth semester as a computer scientist/applied mathematician. My older son David, a journalism student in the Park School at Ithaca College, and I still have two weeks before classes start. Lying back on the sofa proofreading and revising the text while the cats sleep by the fire, it seems to me that academic life, despite its many frustrations, sure beats working for a living.

Rick Durrett

Several figures included here come from other sources and are reprinted with permission of the publisher in parentheses. Figures 3.6, 3.7, and 3.8, from Hudson (2001), Figures 6.3 and 6.4 from Hudson and Kaplan (1988), and Figure 6.6 from Hudson and Kaplan (1995) (Genetics Society of America). Figure 8.7 from Lynch and Conrey (2000) (AAAS). Figure 4.3 from Cann, Stoneking and Wilson (1987) (Nature Publishing Company).

Contents

1

Basic Models

"All models are wrong, but some are useful." George Box

1.1 ATGCs of life

Before we can discuss modeling the evolution of DNA sequences, the reader needs a basic knowledge of the object being modeled. Biologists should skip this very rapid introduction, the purpose of which is to introduce some of the terminology used in later discussions. Mathematicians should concentrate here on the description of the genetic code and the notion of recombination. An important subliminal message is that DNA sequences are not long words randomly chosen from a four-letter alphabet; chemistry plays an important role as well.

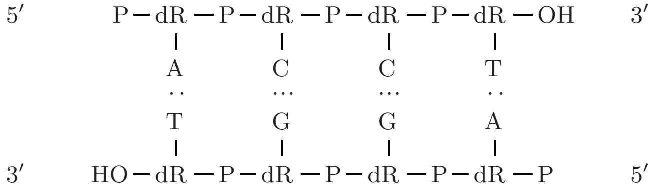

5′
P – dR – P – dR – P – dR – P – dR – OH
3′
A
C
C
T
T
G
G
A
3′
HO – dR – P – dR – P – dR – P – dR – P
5′

Fig. 1.1. Structure of DNA.

The hereditary information of most living organisms is carried by deoxyribonucleic acid (DNA) molecules. DNA usually consists of two complementary

R. Durrett, *Probability Models for DNA Sequence Evolution*,
DOI: 10.1007/978-0-387-78168-6_1, © Springer Science+Business Media, LLC 2008

chains twisted around each other to form a double helix. As drawn in the figure, each chain is a linear sequence of four *nucleotides*: adenine (A), guanine (G), cytosine (C), and thymine (T). Adenine pairs with thymine by means of two hydrogen bonds, while cytosine pairs with guanine by means of three hydrogen bonds. The $A = T$ bond is weaker than the $C \equiv G$ one and separates more easily. The backbone of the DNA molecule consists of sugars (deoxyribose, dR) and phosphates (P) and is oriented. There is a phosphoryl radical (P) on one end (the $5'$ end) and a hydroxyl (OH) on the other ($3'$ end). By convention, DNA sequences are written in the order in which they are transcribed from the $5'$ to the $3'$ end.

The structure of DNA guarantees that the overall frequencies of A and T are equal and that the frequencies of C and G are equal. Indeed, this observation was one of the clues to the structure of DNA. If DNA sequences were constructed by rolling a four-sided die, then all four nucleotides (which are also called *base pairs*) would have a frequency near 1/4, but they do not. If one examines the 12 million nucleotide sequence of the yeast genome, which consists of the sequence of one strand of each of its 16 chromosomes, then the frequencies of the four nucleotides are

$$A = 0.3090 \qquad T = 0.3078 \qquad C = 0.1917 \qquad G = 0.1913$$

Watson and Crick (1953a), in their first report on the structure of DNA, wrote: “It has not escaped our attention that the specific [nucleotide base] pairing we have postulated immediately suggests a possible copying mechanism of the genetic material.” Later that year at a Cold Spring Harbor meeting, Watson and Crick (1953b) continued: “We should like to propose . . . that the specificity of DNA replication is accomplished without recourse to specific protein synthesis and that each of our complimentary DNA chains serves as a template or mould for the formation onto itself of a new companion chain.” This picture turned out to be correct. When DNA is ready to multiply, its two strands pull apart, along each one a new strand forms in the only possible way, and we wind up with two copies of the original. The precise details of the replication process are somewhat complicated, but are not important for our study.

Much of the sequence of the 3 billion nucleotides that make up the human genome apparently serves no function, but embedded in this long string are about 30,000 protein-coding genes. These genes are *transcribed* into ribonucleic acid (RNA), so-called messenger RNA (mRNA), which subsequently is *translated* into proteins. RNA is usually a single-stranded molecule and differs from DNA by having ribose as its backbone sugar and by using the nucleotide uracil (U) in place of thymine (T).

Amino acids are the basic structural units of proteins. All proteins in all organisms, from bacteria to humans, are constructed from 20 amino acids. The next table lists them along with their three-letter and one-letter abbreviations.

Ala	A	Alanine	Leu	L	Leucine
Arg	R	Arginine	Lys	K	Lysine
Asn	N	Asparagine	Met	M	Methionine
Asp	D	Aspartic acid	Phe	F	Phenylalanine
Cys	C	Cysteine	Pro	P	Proline
Gly	G	Glycine	Ser	S	Serine
Glu	E	Glutamic acid	Thr	T	Threonine
Gln	Q	Glutamine	Trp	W	Tryptophan
His	H	Histidine	Tyr	Y	Tyrosine
Ile	I	Isoleucine	Val	V	Valine

Amino acids are coded by triplets of adjacent nucleotides called *codons*. Of the 64 possible triplets, 61 code for amino acids, while 3 are stop codons, which terminate transcription. The correspondence between triplets of RNA nucleotides and amino acids is given by the following table. The first letter of the codon is given on the left edge, the second on the top, and the third on the right. For example, CAU codes for Histidine.

	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	”	”	”	”	C
	Leu	”	Stop	Stop	A
	”	”	”	Trp	G
C	Leu	Pro	His	Arg	U
	”	”	”	”	C
	”	”	Gln	”	A
	”	”	”	”	G
A	Ile	Thr	Asn	Ser	U
	”	”	”	”	C
	”	”	Lys	Arg	A
	Met	”	”	”	G
G	Val	Ala	Asp	Gly	U
	”	”	”	”	C
	”	”	Glu	”	A
	”	”	”	”	G

Note that in 8 of 16 cases, the first two nucleotides determine the amino acid, so a mutation that changes the third base does not change the amino acid that is coded for. Mutations that do not change the amino acid are called *synonymous substitutions*; the others are *nonsynonymous*. For example, a change at the second position always changes the amino acid coded for, except for $UAA \to UGA$, which are both stop codons.

In DNA, adenine and guanine are *purines* while cytosine and thymine are *pyrimidines*. A substitution that preserves the type is called a *transition*; the others are called *transversions*. As we will see later in this chapter, transitions occur more frequently than transversions.

Most of the genes in our bodies reside on DNA in the nucleus of our cells and are organized into chromosomes. Lower organisms such as bacteria are *haploid.* They have one copy of their genetic material. Most higher organisms are *diploid* (i.e., have two copies). However, some plants are *tetraploid* (four copies), *hexaploid* (six copies, e.g., wheat), or *polyploid* (many copies, e.g., sorghum, which has more than 100 chromosomes of 8 basic types). Sex chromosomes in diploids are an exception to the two-copy rule. In humans, females have two X chromosomes, while males have one X and one Y. In birds, males have two Z chromosomes, while females have one Z and one W.

When haploid individuals reproduce, there is one parent that passes copies of its genetic material to its offspring. When diploid individuals reproduce, there are two parents, each of which contributes one of each of its pairs of chromosomes. Actually, one parent's contribution may be a combination of its two chromosomes, since *homologous* pairs (e.g., the two copies of human chromosome 14) undergo recombination, a reciprocal exchange of genetic material that may be diagrammed as follows:

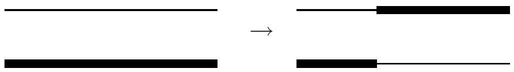


Fig. 1.2. Recombination between homologous chromosomes.

As we will see in Chapter 3, recombination will greatly complicate our analysis. Two cases with no recombination are the Y chromosome, which except for a small region near the tip does not recombine, and the mitochondrial DNA (mtDNA), a circular double-stranded molecule about 16,500 base pairs in length that exist in multiple identical copies outside the nucleus and are inherited from the maternal parent. mtDNA, first sequenced by Anderson et al. (1981), contains genes that code for 13 proteins, 22 tRNA genes, and 2 rRNA genes. It is known that nucleotide substitutions in mtDNA occur at about 10 times the rate for nuclear genes. One important part of the molecule is the control region (sometimes referred to as the D loop), which is about 1,100 base pairs in length and contains promoters for transcription and the origin of replication for one of the DNA strands. It has received particular attention since it has an even higher mutation rate, perhaps an order of magnitude larger than the rest of the mtDNA.

These definitions should be enough to get the reader started. We will give more explanations as the need arises. Readers who find our explanations of the background insufficient should read the *Cartoon Guide to Genetics* by Gonick and Wheelis (1991) or the first chapter of Li's (1997) *Molecular Evolution.*

1.2 Wright-Fisher model

We will begin by considering a genetic locus with two alleles A and a that have the same fitness in a diploid population of constant size N with nonoverlapping generations that undergoes random mating. The first thing we have to do is to explain the terms in the previous sentence.

A *genetic locus* is just a location in the genome of an organism. A common example is the sequence of nucleotides that makes up a gene.

The two *alleles, A and a,* could be the "wrinkled" and "round" types of peas in Mendel's experiments. More abstractly, alleles are just different versions of the genetic information encoded at the locus.

The *fitness* of an individual is a measure of the individual's ability to survive and to produce offspring. Here we consider the case of *neutral evolution* in which the mutation changes the DNA sequence but this does not change the fitness.

Diploid individuals have two copies of their genetic material in each cell. In general, we will treat the N individuals as $2N$ copies of the locus and not bother to pair the copies to make individuals. Note: It may be tempting to set $M = 2N$ and reduce to the case of M haploid individuals, but that makes it harder to compare with formulas in the literature.

To explain the terms *nonoverlapping generations* and *random mating*, we use a picture.

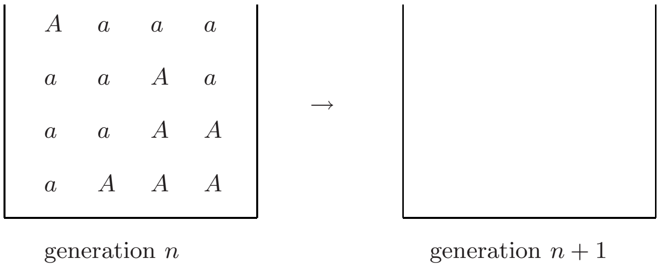


Fig. 1.3. Wright-Fisher model.

In words, we can represent the state of the population in generation n by an "urn" that contains $2N$ balls: i with A's on them and $2N - i$ with a's. Then, to build up the $(n + 1)$th generation, we choose at random from the urn $2N$ times with replacement.

Let X_n be the number of A's in generation n. It is easy to see that X_n is a Markov chain, i.e., given the present state, the rest of the past is irrelevant for predicting the future. Remembering the definition of the binomial distribution, it is easy to see that the probability there are j A's at time $n+1$ when there are i A's in the urn at time n is

$$p(i,j) = \binom{2N}{j} p_i^j (1-p_i)^{2N-j} \tag{1.1}$$

Here $p_i = i/2N$ is the probability of drawing an A on one trial when there are i in the urn, and the *binomial coefficient*

$$\binom{2N}{j} = \frac{(2N)!}{j!(2N-j)!}$$

is the number of ways of choosing j things out of $2N$, where $j! = 1 \cdot 2 \cdots j$ is "j factorial."

Fixation probability

The long-time behavior of the Wright-Fisher model is not very exciting. Since we are, for the moment, ignoring mutation, eventually the number of A's in the population, X_n, will become 0, indicating the loss of the A allele, or $2N$, indicating the loss of a. Once one allele is lost from the population, it never returns, so the states 0 and $2N$ are *absorbing states* for X_n. That is, once the chain enters one of these states, it can never leave. Let

$$\tau = \min\{n : X_n = 0 \text{ or } X_n = 2N\}$$

be the *fixation time*; that is, the first time that the population consists of all a's or all A's.

We use P_i to denote the probability distribution of the process X_n starting from $X_0 = i$, and E_i to denote expected value with respect to P_i.

Theorem 1.1. *In the Wright-Fisher model, the probability of fixation in the all A's state,*

$$P_i(X_\tau = 2N) = \frac{i}{2N} \tag{1.2}$$

Proof. Since the number of individuals is finite, and it is always possible to draw either all A's or all a's, fixation will eventually occur. Let X_n be the number of A's at time n. Since the mean of the binomial in (1.1) is $2Np$, it follows that

$$E(X_{n+1}|X_n = i) = 2N \cdot \left(\frac{i}{2N}\right) = i = X_n \tag{1.3}$$

Taking expected value, we have $EX_{n+1} = EX_n$. In words, the average value of X_n stays constant in time.

Intuitively, the last property implies

$$i = E_i X_\tau = 2NP_i(X_\tau = 2N)$$

and this gives the desired formula. To prove this, we note that since $X_n = X_\tau$ when $n > \tau$,

$$i = E_i X_n = E_i(X_\tau; \tau \le n) + E_i(X_n; \tau > n)$$

where $E(X; A)$ is short for the expected value of X over the set A. Now let $n \to \infty$ and use the fact that $|X_n| \le 2N$ to conclude that the first term converges to $E_i X_\tau$ and the second to 0. □

From (1.2) we get a famous result of Kimura:

Theorem 1.2. *Under the Wright-Fisher model, the rate of fixation of neutral mutations in a population of size N is the mutation rate μ.*

Proof. To see this note that mutations occur to some individual in the population at rate $2N\mu$ and go to fixation with probability $1/2N$.

Heterozygosity

To get an idea of how long fixation takes to occur, we will examine the *heterozygosity*, which we define here to be the probability that two copies of the locus chosen (without replacement) at time n are different:

$$H_n^o = \frac{2X_n(2N - X_n)}{2N(2N-1)}$$

Theorem 1.3. *Let $h(n) = EH_n^o$ be the average value of the heterozygosity at time n. In the Wright-Fisher model*

$$h(n) = \left(1 - \frac{1}{2N}\right)^n \cdot h(0) \tag{1.4}$$

Proof. It is convenient to number the $2N$ copies of the locus $1, 2, \dots 2N$ and refer to them as individuals. Suppose we pick two individuals numbered $x_1(0)$ and $x_2(0)$ at time n. Each individual $x_i(0)$ is a descendant of some individual $x_i(1)$ at time $n-1$, who is a descendant of $x_i(2)$ at time $n-2$, etc. $x_i(m)$, $0 \le m \le n$ describes the lineage of $x_i(0)$, i.e., its ancestors working backwards in time.

If $x_1(m) = x_2(m)$, then we will have $x_1(\ell) = x_2(\ell)$ for $m < \ell \le n$. If $x_1(m) \ne x_2(m)$, then the two choices of parents are made independently, so $x_1(m+1) \ne x_2(m+1)$ with probability $1-(1/2N)$. In order for $x_1(n) \ne x_2(n)$, different parents must be chosen at all times $1 \le m \le n$, an event with probability $(1 - 1/2N)^n$. When the two lineages avoid each other, $x_1(n)$ and $x_2(n)$ are two individuals chosen at random from the population at time 0, so the probability that they are different is $H_0^o = h(0)$. □

A minor detail. If we choose with replacement above, then the statistic is

$$H_n = \frac{2X_n(2N - X_n)}{(2N)^2} = \frac{2N-1}{2N} H_n^o$$

and we again have $EH_n = (1 - 1/2N)^n \cdot H_0$. This version of the statistic is more commonly used, but is not very nice for the proof given above.

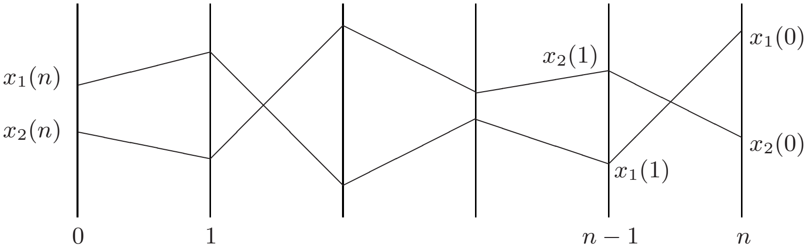


Fig. 1.4. A pair of genealogies.

1.2.1 The coalescent

When x is small, we have $(1-x) \approx e^{-x}$. Thus, when N is large, (1.4) can be written as

$$h(n) \approx e^{-n/(2N)} h(0)$$

If we sample k individuals, then the probability that two will pick the same parent from the previous generation is

$$\approx \frac{k(k-1)}{2} \cdot \frac{1}{2N}$$

where the first factor gives the number of ways of picking two of the k individuals and the second the probability they will choose the same parent. Here we are ignoring the probability that two different pairs choose the same parents on one step or that three individuals will all choose the same parent, events of probability of order $1/N^2$.

Theorem 1.4. *When measured in units of $2N$ generation, the amount of time during which there are k lineages, t_k, has approximately an exponential distribution with mean $2/k(k-1)$.*

Proof. By the reasoning used above, the probability that the k lineages remain distinct for the first n generations is (when the population size N is large)

$$\approx \left(1 - \frac{k(k-1)}{2} \cdot \frac{1}{2N}\right)^n \approx \exp\left(-\frac{k(k-1)}{2} \cdot \frac{n}{2N}\right)$$

Recalling that the exponential distribution with rate λ is defined by

$$P(T > t) = e^{-\lambda t}$$

and has mean $1/\lambda$, we see that if we let the population size $N \to \infty$ and express time in terms of $2N$ generations, that is, we let $t = n/(2N)$, then the time to the first collision converges to an exponential distribution with mean $2/k(k-1)$. Using terminology from the theory of continuous-time Markov chains, k lineages coalesce to $k-1$ at rate $k(k-1)/2$. Since this reasoning applies at any time at which there are k lineages, the desired result follows. □

The limit of the genealogies described in Theorem 1.4 is called the *coalescent*. Letting T_j be the first time that there are j lineages, we can draw a picture of what happens to the lineages as we work backwards in time:

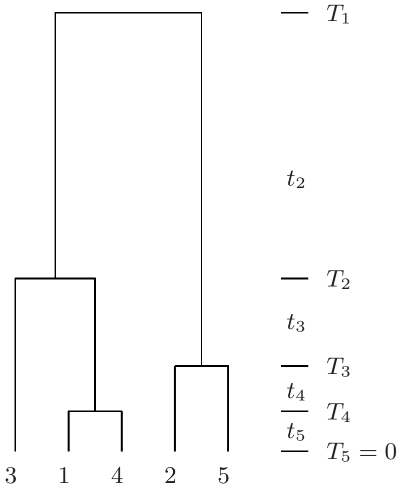


Fig. 1.5. A realization of the coalescent for a sample of size 5.

For simplicity, we do not depict how the lineages move around in the set before they collide, but only indicate when the coalescences occur. To give the reader some idea of the relative sizes of the coalescent times, we have made the t_k proportional to their expected values, which in this case are

$$Et_2 = 1, \quad Et_3 = 1/3, \quad Et_4 = 1/6, \quad Et_5 = 1/10$$

T_1 is the time of the *most recent common ancestor (MRCA)* of the sample. For a sample of size n, $T_1 = t_n + \cdots + t_2$, so the mean

$$ET_1 = \sum_{k=2}^{n} \frac{2}{k(k-1)} = 2 \sum_{k=2}^{n} \left(\frac{1}{k-1} - \frac{1}{k} \right) = 2 \cdot \left(1 - \frac{1}{n} \right)$$

This quantity converges to 2 as the sample size $n \to \infty$, but the time, t_2, at which there are only two lineages has $Et_2 = 1$, so the expected amount of time spent waiting for the last coalescence is always at least half of the expected total coalescence time.

Simulating the coalescent

It is fairly straightforward to translate the description above into a simulation algorithm, but for later purposes it is useful to label the internal nodes of the tree. The following picture should help explain the procedure

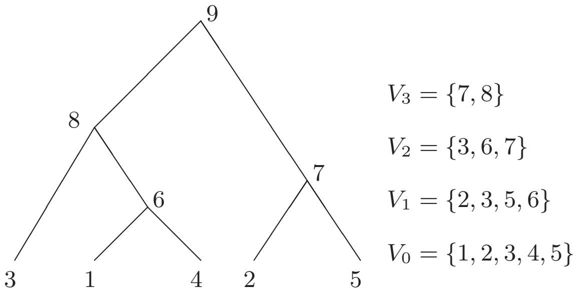


Fig. 1.6. Notation for the coalescent simulation algorithm.

For a sample of size n, we begin with $V_0 = \{1, 2, \dots n\}$ and $T_n = 0$.
For $k = 0, 1, \dots n-2$ do

- Pick two numbers i_k and j_k from V_k.
- Let $V_{k+1} = V_k - \{i_k, j_k\} \cup \{n+k+1\}$.
- In the tree connect $i_k \to n+k+1$ and $j_k \to n+k+1$.
- Let t_{n-k} be an independent exponential with mean $\binom{n-k}{2}^{-1}$.
- Let $T_{n-k-1} = T_{n-k} + t_{n-k}$.

To implement this in a computer, one can let $t_{n-k} = \binom{n-k}{2}^{-1} \log(1/U_k)$, where the U_k are independent uniform$(0,1)$. From the construction it should be clear that the sequence of coalescence events is independent of the sequence of times of interevent times $t_n, \dots t_2$.

In the next two sections, we will introduce mutations. To do this in the computer, it is convenient to define the ancestor function in the third step of the algorithm above so that anc$[i_k] = n+k+1$ and anc$[j_k] = n+k+1$. For example, anc$[2] = 7$ and anc$[5] = 7$. One can then label the branches by the smaller number $1 \le i \le 2n-2$ on the lower end and, if mutations occur at rate μ per generation and $\theta = 4N\mu$, introduce a Poisson number of mutations on branch i with mean

$$\frac{\theta}{2} \cdot (T_{\mathrm{anc}[i]} - T_i)$$

The reason for the definition of θ will become clear in Section 1.4. Before we move on, the reader should note that we first generate the genealogy and then introduce mutations.

1.2.2 Shape of the genealogical tree

The state of the coalescent at any time can then be represented as a *partition*, $A_1, \dots A_m$, of $\{1, 2, \dots n\}$. That is, $\cup_{i=1}^m A_i = \{1, 2, \dots n\}$, and if $i \neq j$ the sets A_i and A_j are disjoint. In words, each A_i consists of one subset of lineages that have coalesced. To explain this notion, we will use the example that appears in the two previous figures. In this case, as we work backwards in time, the partitions are

T_1	$\{1,2,3,4,5\}$				
T_2	$\{1,3,4\}$	$\{2,5\}$			
T_3	$\{1,4\}$	$\{2,5\}$	$\{3\}$		
T_4	$\{1,4\}$	$\{2\}$	$\{3\}$	$\{5\}$	
time 0	$\{1\}$	$\{2\}$	$\{3\}$	$\{4\}$	$\{5\}$

Initially, the partition consists of five singletons since there has been no coalescence. After 1 and 4 coalesce at time T_4, they appear in the same set. Then 2 and 5 coalesce at time T_3, etc. Finally, at time T_1 we end up with all the labels in one set.

Let $\mathcal{E}_n$ be the collection of partitions of $\{1, 2, \dots n\}$. If $\xi \in \mathcal{E}_n$, let $|\xi|$ be the number of sets that make up ξ, i.e., the number of lineages that remain in the coalescent. If, for example, $\xi = \{\{1,4\}, \{2,5\}, \{3\}\}$, then $|\xi| = 3$. Let ξ_i^n, $i = n, n-1, \dots 1$ be the partition of $\{1, 2, \dots n\}$ at time T_i, the first time there are i lineages. Kingman (1982a) has shown

Theorem 1.5. *If ξ is a partition of $\{1, 2, \dots n\}$ with $|\xi| = i$, then*

$$P(\xi_i^n = \xi) = c_{n,i}\, w(\xi)$$

Here the weight $w(\xi) = \lambda_1! \cdots \lambda_i!$, where $\lambda_1, \dots \lambda_i$ are the sizes of the i sets in the partition and the constant

$$c_{n,i} = \frac{i!}{n!} \cdot \frac{(n-i)!(i-1)!}{(n-1)!}$$

is chosen to make the sum of the probabilities equal to 1.

The proof of Theorem 1.6 will give some insight into the form of the constant. The weights $w(\xi)$ favor partitions that are uneven. For example, if $n = 9$ and $i = 3$, the weights based on the sizes of the sets in the partition are as follows:

3-3-3	4-3-2	5-2-2	4-4-1	5-3-1	6-2-1	7-1-1
216	288	480	576	720	1440	5040

Proof. We proceed by induction working backwards from $i = n$. When $i = n$, the partition is always $\{1\}, \dots \{n\}$, all the $\lambda_i = 1$, and $c_{n,n} = 1$ (by definition, $0! = 1$). To begin the induction step now, write $\xi < \eta$ (and say ξ is finer than η) if $|\xi| = |\eta| + 1$ and η is obtained by combining two of the sets in ξ. For example, we might have

$$\xi = \{\{1,4\},\ \{2,5\},\ \{3\}\} \quad \text{and} \quad \eta = \{\{1,3,4\},\ \{2,5\}\}$$

When $\xi < \eta$ and $|\xi| = i$, there is exactly one of the $\binom{i}{2}$ coalescence events that will turn ξ into η, so

$$P(\xi^n_{i-1} = \eta | \xi^n_i = \xi) = \begin{cases} \frac{2}{i(i-1)} & \text{if } \xi < \eta \\ 0 & \text{otherwise} \end{cases} \tag{1.5}$$

and we have

$$P(\xi^n_{i-1} = \eta) = \frac{2}{i(i-1)} \sum_{\xi < \eta} P(\xi^n_i = \xi) \tag{1.6}$$

If $\lambda_1, \dots \lambda_{i-1}$ are the sizes of the sets in η, then for some ℓ with $1 \le \ell \le i-1$ and some ν with $1 \le \nu < \lambda_\ell$, the sets in ξ have sizes

$$\lambda_1, \dots \lambda_{\ell-1}, \nu, \lambda_\ell - \nu, \lambda_{\ell+1}, \dots \lambda_{i-1}$$

Using the induction hypothesis, the right-hand side of (1.6) is

$$= \frac{2}{i(i-1)} \sum_{\ell=1}^{i-1} \sum_{\nu=1}^{\lambda_\ell - 1} c_{n,i}\, w_{\ell,\nu} \binom{\lambda_\ell}{\nu} \cdot \frac{1}{2}$$

where the weight

$$w_{\ell,\nu} = \lambda_1! \cdots \lambda_{\ell-1}!\, \nu!\, (\lambda_\ell - \nu)!\, \lambda_{\ell+1}! \cdots \lambda_{i-1}!$$

and $\binom{\lambda_\ell}{\nu} \cdot \frac{1}{2}$ gives the number of ways of picking $\xi < \eta$ with the ℓth set in η subdivided into two pieces of size ν and $\lambda_\ell - \nu$. (We pick ν of the elements to form a new set but realize that we will generate the same choice again when we pick the $\lambda_\ell - \nu$ members of the complement.)

It is easy to see that $w_{\ell,\nu}\binom{\lambda_\ell}{\nu} = w(\eta)$ so the sum above is

$$= w(\eta) \frac{c_{n,i}}{i(i-1)} \sum_{\ell=1}^{i-1} \sum_{\nu=1}^{\lambda_\ell - 1} 1$$

The double sum $= \sum_{\ell=1}^{i-1} (\lambda_\ell - 1) = n - (i-1)$. The last detail to check is that

$$\frac{c_{n,i}}{i(i-1)} \cdot (n - i + 1) = c_{n,i-1} \quad \text{or} \quad \frac{c_{n,i}}{c_{n,i-1}} = \frac{i(i-1)}{n-i+1} \tag{1.7}$$

but this is clear from the definition. □

To write the partition ξ_i^n, it is natural, as we have done in the example above, to order the sets so that $\xi_{i,1}^n$ is the set containing 1, $\xi_{i,2}^n$ contains the smallest number not in $\xi_{i,1}^n$, etc. However, to compute the distribution of the sizes of sets in the coalescent, it is useful to put the sets in the partition into a randomly chosen order.

Theorem 1.6. *Let π be a randomly chosen permutation of $\{1, 2, \dots i\}$ and let $\lambda_j = |\xi_{i,\pi(j)}^n|$ be the size of the jth set in ξ_i^n when they are rearranged according to π. $(\lambda_1, \lambda_2, \dots \lambda_i)$ is uniformly distributed over the vectors of positive integers that add to n.*

Tajima (1983) proved this in the case $i = 2$. In words, if we pick one of the two sets in ξ_2^n at random, its size is uniformly distributed on $1, 2, \dots n - 1$.

Proof. If we randomly order the sets in ξ, then each ordered arrangement has probability $c_{n,i}w(\xi)/i!$. If we only retain information about the sizes, then by considering the number of collections of sets that can give rise to the vector $(\lambda_1, \dots \lambda_i)$, we see that it has probability:

$$\frac{c_{n,i}w(\xi)}{i!} \cdot \frac{n!}{\lambda_1!\lambda_2! \cdots \lambda_i!} = \frac{(n-i)!(i-1)!}{(n-1)!} = 1 \Big/ \binom{n-1}{i-1}$$

Since the final quantity depends only on n and i and not on the actual vector, we have shown that the distribution is uniform. To see that the denominator of the last fraction gives the number of vectors of positive integers of length i that add up to n, imagine n balls separated into i groups by $i - 1$ pieces of cardboard. For example, if $n = 10$ and $i = 4$, we might have

$$O\,O\,O|O|O\,O\,O\,O|O\,O$$

Our $i - 1$ pieces of cardboard can go in any of the $n - 1$ spaces between balls, so there are $\binom{n-1}{i-1}$ possible vectors $(j_1, \dots, j_i)$ of positive integers that add up to n. □

As a consequence of Theorem 1.6, we get the following amusing fact.

Theorem 1.7. *The probability that the most recent common ancestor of a sample of size n is the same as that of the population converges to $(n-1)/(n+1)$ as the population size tends to ∞.*

When $n = 10$, this is 9/11.

Proof. Consider the first split in the coalescent tree of the population. Let X be the limiting proportion of lineages in the left half of the tree, and recall that X is uniformly distributed on $(0, 1)$. In order for the MRCA of the sample to come before that of the population either all of the n lineages must be in the left half or all n in the right half. Thus, the probability the MRCAs coincide is

$$1 - \int_0^1 x^n + (1-x)^n \, dx = 1 - \frac{2}{n+1}$$

and this gives the desired result. □

Our final result in this section gives a dynamic look at the coalescent process running backwards.

Theorem 1.8. *To construct the partition ξ_i^n from $\xi_{i-1}^n = \{A_1, \dots A_{i-1}\}$, we pick a set at random, picking A_j with probability $(\lambda_j - 1)/(n-i-1)$, where $\lambda_j = |A_j|$, and then split A_j into two sets with sizes k and $\lambda_j - k$, where k is uniform on $1, 2, \dots \lambda_j - 1$.*

Proof. By elementary properties of conditional probability,

$$P(\xi_i^n = \xi | \xi_{i-1}^n = \eta) = \frac{P(\xi_{i-1}^n = \eta | \xi_i^n = \xi) P(\xi_i^n = \xi)}{P(\xi_{i-1}^n = \eta)}$$

Suppose $\xi < \eta$ are partitions of the correct sizes, and ξ is obtained from η by splitting A_j into two sets with sizes k and $\lambda_j - k$. It follows from Theorem 1.5 and (1.5) that

$$\frac{P(\xi_i^n = \xi)}{P(\xi_{i-1}^n = \eta)} = \frac{c_{n,i}}{c_{n,i-1}} \cdot \frac{k!(\lambda_j - k)!}{\lambda_j!} = \frac{i(i-1)}{n-i+1} \cdot \frac{k!(\lambda_j - k)!}{\lambda_j!}$$

Using (1.7), now we have

$$P(\xi_i^n = \xi | \xi_{i-1}^n = \eta) = \frac{1}{n-i+1} \cdot \frac{k!(\lambda_j - k)!}{\lambda_j!} \cdot 2$$

The first factor corresponds to picking A_j with probability $(\lambda_j - 1)/(n-i-1)$ and then picking k with probability $1/(\lambda_j - 1)$. Getting the correct division of A_j to produce ξ has probability $1/\binom{\lambda_j}{k}$. The final 2 takes into account the fact that we can also generate ξ by choosing $\lambda_j - k$ instead of k. □

1.3 Infinite alleles model

In this section, we will consider the *infinite alleles model.* As the name should suggest, we assume that there are so many alleles that each mutation is always to a new type never seen before. To explain the reason for this assumption, Kimura (1971) argued that if a gene consists of 500 nucleotides, the number of possible DNA sequences is

$$4^{500} = 10^{500 \log 4 / \log 10} = 10^{301}$$

For any of these, there are $3 \cdot 500 = 1500$ sequences that can be reached by single base changes, so the chance of returning where one began in two mutations is 1/1500 (assuming an equal probability for all replacements). Thus, the total number of possible alleles is essentially infinite.

The infinite alleles model arose at a time when one had to use indirect methods to infer diferences between individuals. For example, Coyne (1976) and Singh, Lewontin, and Felton (1976) studied *Drosophila* by performing electrophoresis under various conditions. Coyne (1976) found 23 alleles in 60 family lines at the xanthine dehydrogenase locus of *Drosophila persimilis* that displayed the following pattern, which we call the *allelic partition*:

$$a_1 = 18,\ a_2 = 3,\ a_4 = 1,\ a_{32} = 1$$

That is, there were 18 unique alleles, 3 alleles had 2 representatives, 1 had 4, and 1 had 32. Singh, Lewontin, and Felton (1976) found 27 alleles in 146 genes from the xanthine dehydrogenase locus of *D. pseudoobscura* with the following pattern:

$$a_1 = 20,\ a_2 = 3,\ a_3 = 7,\ a_5 = 2,\ a_6 = 2,\ a_8 = 1,\ a_{11} = 1,\ a_{68} = 1$$

The infinite alleles model is also relevant to DNA sequence data when there is no recombination. Underhill et al. (1997) studied 718 Y chromosomes. They found 22 nucleotides that were *polymorphic* (i.e., not the same in all of the individuals). The sequence of nucleotides at these variable positions gives the *haplotype* of the individual. In the sample, there were 20 distinct haplotypes. The sequences can be arranged in a tree in which no mutation occurs more than once, so it is reasonable to assume that the haplotypes follow the infinite alleles model. The allelic partition has

$$\begin{aligned} a_1 = 7,\ a_2 = a_3 = a_5 = a_6 = a_8 = a_9 = a_{26} = a_{36} = a_{37} = 1, \\ a_{82} = 2, a_{149} = 1, a_{266} = 1. \end{aligned}$$

After looking at the data, the first obvious question is: What do we expect to see? The answer to this question is given by *Ewens' sampling formula.* This section is devoted to the derivation of the formula and the description of several perspectives from which one can approach it. At the end of this section, we will lapse into a mathematical daydream about the structure of a randomly chosen permutation.

1.3.1 Hoppe's urn, Ewens' sampling formula

The genealogical process associated with the infinite alleles version of the Wright-Fisher model is a coalescent with killing. When there are k lineages, coalescence and mutation occur on each step with probability

$$\frac{k(k-1)}{2} \cdot \frac{1}{2N}$$

as before, but now killing of one of the lineages occurs with probability $k\mu$ because if a mutation is encountered, we know the genetic state of that individual and all of its descendants in the sample. Speeding up the system by running time at rate $2N$, the rates become $k(k-1)/2$ and $k\theta/2$, where $\theta = 4N\mu$.

Turning the coalescent with killing around backwards leads to *Hoppe's (1984) urn model.* This urn contains a black ball with mass θ and various colored balls with mass 1. At each time, a ball is selected at random with a probability proportional to its mass. If a colored ball is drawn, that ball and another of its color are returned to the urn. If the black ball is chosen, it is returned to the urn with a ball of a new color that has mass 1. The choice of the black ball corresponds to a new mutation and the choice of a colored ball corresponds to a coalescence event. A simulation should help explain the definition. Here a black dot indicates that a new color was added at that time step.

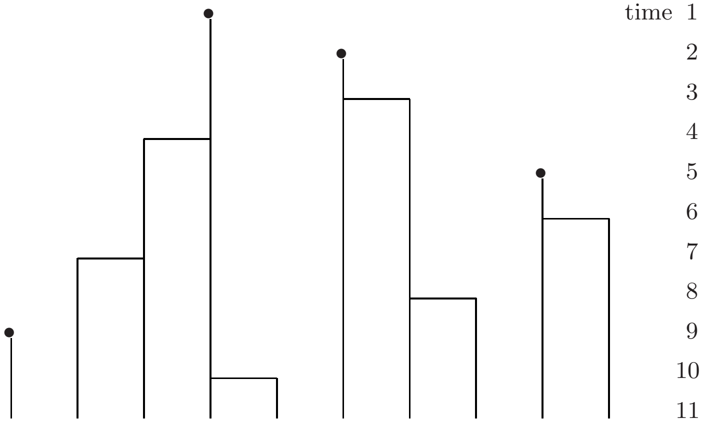


Fig. 1.7. A realization of Hoppe's urn.

As we go backwards from time $k+1$ to time k in Hoppe's urn, we encounter a mutation with probability $\theta/(\theta+k)$ and have a coalescence with probability $k/(\theta+k)$. Since in the coalescent there are $k+1$ lineages that are each exposed to mutations at total rate $k\theta/2$ and collisions occur at rate $(k+1)k/2$, this is the correct ratio. Since by symmetry all of the coalescence events have equal probability, it follows that

Theorem 1.9. *The genealogical relationship between k lineages in the coalescent with killing can be simulated by running Hoppe's urn for k time steps.*

This observation is useful in computing properties of population samples under the infinite alleles model. To illustrate this, let K_n be the random variable that counts the number of different alleles found in a sample of size n. Here and throughout the book, log is the "natural logarithm" with base e.

Theorem 1.10 (Watterson (1975)). *For fixed θ, as the sample size $n \to \infty$,*

$$EK_n \sim \theta \log n \quad \textit{and} \quad \textit{var}\,(K_n) \sim \theta \log n$$

where $a_n \sim b_n$ means that $a_n/b_n \to 1$ as $n \to \infty$. In addition, the central limit theorem holds. That is, if χ has the standard normal distribution, then

$$P\left(\frac{K_n - EK_n}{\sqrt{\textit{var}\,(K_n)}} \le x \right) \to P(\chi \le x)$$

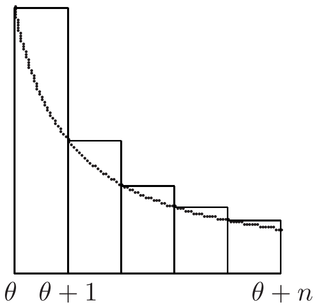
Proof. Let $\eta_i = 1$ if the ith ball added to Hoppe's urn is a new type and 0 otherwise. It is clear from the definition of the urn scheme that $K_n = \eta_1 + \cdots + \eta_n$ and $\eta_1, \ldots \eta_n$ are independent with

$$P(\eta_i = 1) = \theta/(\theta + i - 1) \tag{1.8}$$

To compute the asymptotic behavior of EK_n, we note that (1.8) implies

$$EK_n = \sum_{i=1}^{n} \frac{\theta}{\theta + i - 1} \tag{1.9}$$

Viewing the right-hand side as a Riemann sum approximating an integral,



it follows that

$$\sum_{i=1}^{n} \frac{1}{\theta + i - 1} \sim \int_{\theta}^{n+\theta} \frac{1}{x}\,dx = \log(n + \theta) - \log(\theta) \sim \log n \tag{1.10}$$

From this, the first result follows. To prove the second, we note that

$$\text{var}\,(K_n) = \sum_{i=1}^{n} \text{var}\,(\eta_i) = \sum_{i=2}^{n} \frac{\theta(i-1)}{(\theta + i - 1)^2} \tag{1.11}$$

As $i \to \infty$, $(i-1)/(\theta + i - 1) \to 1$, so

$$\text{var}\,(K_n) \sim \sum_{i=2}^{n} \frac{\theta}{\theta + i - 1} \sim \theta \log n$$

by the reasoning for EK_n. Since the η_i are independent, the final claim follows from the triangular array form of the central limit theorem. See, for example, (4.5) in Chapter 2 of Durrett (2005). □

An immediate consequence of Theorem 1.10 is that $K_n/\log n$ is an asymptotically normal estimator of the scaled mutation rate θ. However, the asymptotic standard deviation of $K_n/\log n$ is quite large, namely of order $1/\sqrt{\log n}$. Thus, if the true $\theta = 1$ and we want to estimate θ with a standard error of 0.1, a sample of size e^{100} is required. Given this depressingly slow rate of convergence, it is natural to ask if there is another way to estimate θ from the data. The answer is NO, however. As we will see in Theorem 1.13 below, K_n is a sufficient statistic. That is, it contains all the information in the sample that is useful for estimating θ.

The last result describes the asymptotic behavior of the number of alleles. The next one, due to Ewens (1972), deals with the entire distribution of the sample under the infinite alleles model.

Theorem 1.11 (Ewens' sampling formula). *Let a_i be the number of alleles present i times in a sample of size n. When the scaled mutation rate is $\theta = 4N\mu$,*

$$P_{\theta,n}(a_1, \ldots a_n) = \frac{n!}{\theta_{(n)}} \prod_{j=1}^{n} \frac{(\theta/j)^{a_j}}{a_j!}$$

where $\theta_{(n)} = \theta(\theta+1)\cdots(\theta+n-1)$.

The formula may look strange at first, but it becomes more familiar if we rewrite it as

$$c_{\theta,n} \prod_{j=1}^{n} e^{-\theta/j} \frac{(\theta/j)^{a_j}}{a_j!}$$

where $c_{\theta,n}$ is a constant that depends on θ and n and guarantees the sum of the probabilities is 1. In words, if we let $Y_1, \ldots Y_n$ be independent Poisson random variables with means $EY_j = \theta/j$, then the allelic partition $(a_1, a_2, \ldots a_n)$ has the same distribution as

$$\left(Y_1, Y_2, \ldots, Y_n \,\middle|\, \sum_m mY_m = n \right)$$

One explanation of this can be found in Theorem 1.19.

Proof. In view of Theorem 1.9, it suffices to show that the distribution of the colors in Hoppe's urn at time n is given by Ewens' sampling formula. We proceed by induction. When $n = 1$, the partition $a_1 = 1$ has probability 1 so