Mathematical Genetics
LMS Durham Symposium July 2004

The Ewens sampling formula and related formulas: combinatorial proofs, extensions to variable population size and applications to ages of alleles

Robert C. GRIFFITHS

University of Oxford

Sabin LESSARD

Université de Montréal

THEORETICAL POPULATION BIOLOGY 3, 87–112 (1972)

The Sampling Theory of Selectively Neutral Alleles*

W. J. EWENS[†]

Department of Zoology, University of Texas at Austin, Austin, Texas, 78712

Received August 17, 1971

DEDICATED TO THE MEMORY OF KEN KOJIMA

In this paper a beginning is made on the sampling theory of neutral alleles. That is, we consider deductive and subsequently inductive questions relating to a sample of genes from a selectively neutral locus. The inductions concern estimation, confidence intervals and hypothesis testing. In particular the test of the hypothesis that the alleles being sampled are indeed selectively neutral will be considered. In view of the large amount of data currently being obtained by electrophoretic methods on allele frequencies and numbers, and the current interest in the possibility of extensive "non-Darwinian" evolution, such a sampling theory seems necessary. However, a large number of unsolved problems in this area remain, a partial listing being given towards the end of this paper.

Mathematical Theory

A number of quantities will be considered in this paper and it is useful to gather together the notations that will be used consistently throughout. We define.

N = number of individuals in the parent (diploid) population (normally unknown)

N_e = effective size of parent population (normally unknown)

u = mutation rate to entirely new alleles (normally unknown)

n = number of individuals in sample (taken in generation t)

K = number of alleles in the population in generation t (an unknown random variable)

k = number of different alleles observed in the sample (a realized value of a random variable)

* Supported by USPHS Grant GM-15769.

[†] Current address: Mathematical Research Center, University of Wisconsin, Madison, WI 53706. Permanent address: Department of Mathematics, La Trobe University, Bundoora 3083, Victoria, Australia.

P-456

Received: November 26, 1980

Appl. Stochastic Proc.

Ref: Lessard

THE COALESCENT

J.F.C. Kingman

Mathematical Institute, University of Oxford, England

The n-coalescent is a continuous-time Markov chain on a finite set of states, which describes the family relationships among a sample of n members drawn from a large haploid population. Its transition probabilities can be calculated from a factorisation of the chain into two independent components, a pure death process and a discrete-time jump chain. For a deeper study, it is useful to construct a more complicated Markov process in which n-coalescents for all values of n are embedded in a natural way.

Keywords: Genetical models, haploid genealogy, random equivalence relations, coupling, exchangeability, Markov process, jump chain.

1. The n-coalescent

For any natural number n, let $\mathcal{E}_n$ denote the finite set of equivalence relations on {1, 2, ..., n}. For R ϵ $\mathcal{E}_n$, denote by $|R|$ the number of equivalence classes of R. A continuous-time

POPULATION GENETICS THEORY - THE PAST AND THE FUTURE

W.J. Ewens
Department of Mathematics
Monash University
Clayton, Victoria 3168
Australia
and
Department of Biology
University of Pennsylvania
Philadelphia, PA 19104
U.S.A.

ABSTRACT. Classical population genetics theory was largely directed towards processes relating to the future. Present theory, by contrast, focuses on the past, and in particular is motivated by the desire to make inferences about the evolutionary processes which have led to the presently observed patterns and nature of genetic variation. There are many connections between the classical prospective theory and the new retrospective theory. However, the retrospective theory introduces ideas not appearing in the classical theory, particularly those concerning the ancestry of the genes in a sample or in the entire population. It also introduces two important new distributions into the scientific literature, namely the Poisson-Dirichlet and the GEM: these are important not only in population genetics, but also in a very wide range in science and mathematics. Some of these are discussed. Population genetics theory has been greatly enriched by the introduction of many new concepts relating to the past evolution of biological populations.

1. Introduction

These notes are based on lectures given to an audience consisting of biologists, statisticians and mathematicians, and they reflect the breadth of interests of the participants. I have preferred to seek out connections and analogies between these disciplines rather than to pursue any topic in depth, to show how questions of interest to geneticists have led to mathematical developments in areas quite different from biology, and how in turn various mathematical developments lead to a more complete understanding of the evolutionary process.

The title does not imply an ambitious attempt to give an overview of population genetics theory. Rather, it can be interpreted in two different and more specific ways. First, it is intended to suggest the view that even the most recent research has its origins in, and often borrows results from, the classical theory. Secondly, it reflects the view (the closing theme in Ewens (1979)) that the direction of interest in population genetics theory is changing from the prospective to the retrospective. The classical theory, aiming to prove the validity of Darwinian evolution as a Mendelian process, was prospective and

S. Lessard (ed.), Mathematical and Statistical Developments of Evolutionary Theory, 177–227.

EWENS' FORMULA

n sampled genes

$\frac{n!}{n_1!\cdots n_k!}$ assignments of k **types**

$\times\frac{1}{b_1!\cdots b_n!}$ if types are **unlabelled** with b_j being the number of types represented j times

$\times n!$ orders of **loss** backward in time by mutation or coalescence, with the convention that one of the genes at random is lost when two coalesce

$\times\frac{\theta}{i(\theta+i-1)}$ if the ith lost is the last of its type

or $\frac{j-1}{i(\theta+i-1)}$ if it is the jth last of its type, for

$i = 1, ..., n$, since we have the rates

•$i\theta/2$ for **mutation** among i genes

•$i(i-1)/2$ for **coalescence** among i genes

•$\theta/2$ for **loss by mutation** of a given gene

•$(j-1)/2$ for **loss by coalescence** of a given gene with one of $j-1$ genes of the same type

The multiplication of all terms gives

$$\frac{n!}{1^{b_1}\cdots n^{b_n}} \cdot \frac{1}{b_1!\cdots b_n!} \cdot \frac{\theta^k}{\theta(\theta+1)\cdots(\theta+n-1)}$$

for the probability of having k types with b_j types represented j times.

WATTERSON'S FORMULA (KINGMAN'S FORMULA IF $\theta = 0$)

n labelled genes traced back to m ancestral genes of types $1, ..., m$

$n_1 \cdots n_m$ possible **ancestral genes**

$\times(n-m)!$ orders of loss of the younger genes

$\times\frac{\theta}{i(\theta+i-1)}$ or $\frac{j-1}{i(\theta+i-1)}$ for $i = m+1, ..., n$,

and this gives

$$\frac{(n-m)!\theta^{k-m}\prod_{l=1}^{m} n_l! \prod_{l=m+1}^{k}(n_l-1)!}{\prod_{i=m+1}^{n} i(\theta+i-1)}$$

for the probability of having n_l given genes of type l for $l = 1, ..., m, m+1, ..., k$.

VARIABLE POPULATION SIZE WITH AGE-ORDERED TYPES TYPE 1 BEING THE OLDEST AND TYPE k THE YOUNGEST

$\frac{n!}{n_1!\cdots n_k!}$ assignments of k types

$\times \prod_{l=1}^{k} n_l \cdot (\sum_{\nu=1}^{l} n_\nu - i_l) \cdots (\sum_{\nu=1}^{l} n_\nu - i_{l+1} + 2)$

$(a_{\mathbf{i}})$ orders of loss with i_l **genes remaining** just before type l is lost by mutation for $l = 1, ..., k$

$\times \frac{\theta}{i[\theta + (i-1)/\lambda(T_i)]}$ or $\frac{(j-1)/\lambda(T_i)}{i[\theta + (i-1)/\lambda(T_i)]}$

given a **time back of loss** T_i when i genes remain and then a **population size** $\lambda(T_i)$ relative to the present size, for $i = 1, ..., n$, which gives

$$\frac{\theta^{k-1}}{\left(\prod_{l=1}^{k} n_l\right)} \sum_{\mathbf{i}} a_{\mathbf{i}} E\left\{\frac{\prod_{l=2}^{k} \lambda(T_{i_l})}{\prod_{i=2}^{n} [\theta\lambda(T_i) + i - 1]}\right\}$$

LADDER INDICES AND HEIGHTS IN AN URN MODEL

The probability of having age-ordered frequencies $n_1, ..., n_k$ in a sample given mutation events when $i_1, ..., i_k$ genes remain is

$$\frac{\prod_{l=2}^{k}(i_l - 1)}{(n-1)!} \cdot \prod_{m=1}^{k} \frac{(\sum_{\nu=1}^{m} n_\nu - i_m)!}{(\sum_{\nu=1}^{m-1} n_\nu - i_m + 1)!}$$

This is also the probability that the **successive maxima** obtained by drawing balls labelled from 1 to n, at random and without replacement, have **increments** $n_1, ..., n_k$ given that the successive maxima occur at the **drawing numbers** $i_1, ..., i_k$.

POPULATION DISTRIBUTION OF AGE-ORDERED FREQUENCIES

As the sample size goes to **infinity**, the relative frequency of the lth oldest type, n_l/n, given $i_1....,i_k$, is distributed as

$$\xi_{l-1}\prod_{\nu=l}^{\infty}(1-\xi_\nu)$$

where all ξ_l are independent and have densities

$$(i_{l+1}-1)(1-z)^{i_{l+1}-2},\ 0<z<1.$$

Moments and Laplace transforms of these frequencies can be deduced and then, in particular, the probability that the **oldest type in the sample** is the **oldest in the population**

$$\sum_{k=1}^{n}(-1)^{k-1}\binom{n}{k}E\left[\prod_{j=2}^{\infty}(1-\frac{k}{k+j-1}\cdot\frac{\theta\lambda(T_j)}{\theta\lambda(T_j)+j-1})\right]$$

All this is **consistent** with the GEM distribution for a population of constant size.

GENEALOGY OF A DERIVED TYPE

A derived type represented r times in a sample of size n is lost by mutation at time back T_{m+1} when there remain $m+1$ genes with probability $p_{m+1}/(\sum_{j=1}^{n-r} p_{j+1})$ where p_{m+1} is

$$\frac{(n-1)!\theta}{r} \cdot \frac{\begin{pmatrix} n-m-1 \\ r-1 \end{pmatrix}}{\begin{pmatrix} n-1 \\ r \end{pmatrix}} \cdot E\left\{\frac{\lambda(T_{m+1})}{\prod_{i=2}^{n}[\theta\lambda(T_i)+i-1]}\right\}$$

Then the distribution of the **age** of the derived type can be studied, by conditioning on $m+1$, under assumptions about the population size and be approximated under conditions such as low θ, large n or small r, and similarly for the **coalescence times** of the genes of the derived type.

CONCLUSION

Combinatorial arguments coupled with the coalescent approach may simplify the proofs of known results and lead to some new ones.