CYCLES, PERMUTATIONS AND THE STRUCTURE OF THE YULE PROCESS WITH IMMIGRATION

Paul JOYCE and Simon TAVARÉ*

Department of Mathematics, University of Utah, Salt Lake City, UT 84112, USA

Received 5 January 1987
Revised 29 April 1987

A novel representation of the linear birth process with immigration is analysed. The state space of the process is the collection of permutations of the integers, written in a particular cyclic form. The stochastic structure of the model is particularly simple to describe. The results serve to explain the combinatorial structure of some sampling formulae that arise in the study of neutral mutations in population genetics.

linear birth process * Yule process

1. Introduction

The linear birth process with immigration may be described briefly as follows: At the points of a Poisson process of rate θ an immigrant enters the population, and initiates a family that evolves according to the laws of a linear birth (Yule) process of rate 1; families initiated at different times evolve independently. Define $\{I(t), t \geq 0\}$ to be the population-size process, $I(t)$ being the total number of individuals alive at time t. It is well known that $I(t)$ has a negative binomial distribution:

$$P(I(t)=n)=\binom{n+\theta-1}{n}\mathrm{e}^{-\theta t}(1-\mathrm{e}^{-t})^n, \quad n=0,1,\ldots; \tag{1.1}$$

cf. Kendall (1949).

In this note, we describe a novel (and more detailed) representation of the birth process with immigration in which we use as our state-space the collection of permutations of positive integers, written as a product of cycles in a particular way. Our description gives a complete history of the families in the process; the order in which the cycles are written corresponds to the order in which the families arose in the population, the sizes of the cycles represent the number of individuals in each family, and the cycles themselves tell how each member of a family is related. This explicit representation in terms of cycles of a permutation serves to explain and simplify much of the combinatorial structure of the process.

* Supported in part by NSF grant DMS 86-08857.

2. The process

An informal description of our process runs as follows: suppose that the population currently has $n-1$ members, labelled $1, 2, \ldots, n-1$ in order of their appearance in the population. The next individual, n, to appear in the population is either an immigrant, in which case n starts a new cycle, or it is an offspring of the existing individual j, where $1 \leqslant j \leqslant n-1$. In this case, the new state is formed by inserting the integer n in the cycle in which j belongs, immediately *to the left* of j. As successive individuals enter the population, we build up the history of the process in the form of permutations written as a product of cycles. For example, if the current state of the process is the permutation (731)(4852)(96), then the oldest family contains individuals 7, 3, 1; 7 is a child of 3, and 3 of 1. The second event that occurred in the process was the immigration of individual 2; individual 4 is the offspring of 2, 5 is also an offspring of 2, and 8 is the offspring of 5. The youngest individual, 9, is an offspring of 6, who is himself an immigrant.

More formally, the state-space of our process can be described as follows. Let Π be a permutation of the first n integers, say $\Pi = i_1 i_2 \cdots i_n$. We write Π as an ordered product $c_1 c_2 \cdots c_k$ of cycles of a particular form. We adopt the convention that the first cycle of Π *starts* with i_1 and *finishes* with the integer 1. If j is the smallest integer not included in the first cycle, the second cycle starts with i_j and ends with j, and so on. For example if $n=9$ and $\Pi = 615784239$ then we write $\Pi = (64721)(583)(9)$ as a product of cycles. (The conventional representation of this particular permutation as a product of cycles would be (16472)(358)(9). Our definition of a cycle makes the stochastic process of interest much easier to describe; in any event the two types of cycle representation are in an obvious one-to-one correspondence so no confusion need arise.) For $n \geqslant 1$, Y_n will denote the collection of permutations of the first n integers written in this *ordered* cycle form. It will be convenient to let (0) denote the permutation of no elements. $Y = \bigcup_{n \geqslant 1} Y_n \bigcup \{(0)\}$ will denote the collection of all permutations.

The birth process with immigration may be represented as a Markov process $\{\Pi(t), t \geqslant 0; \Pi(0) = (0)\}$ on Y which has transition rates $(q_{\pi\eta}, \pi \neq \eta)$ determined as follows: if $\pi = (0)$, then

$$q_{\pi\eta} = \begin{cases} \theta & \text{if } \eta = (1), \\ 0 & \text{otherwise.} \end{cases} \tag{2.1a}$$

If $\pi \in Y_{n-1}$, $n > 1$, and π has the form $\pi = c_1 \cdots c_i \cdots c_k$, then

$$q_{\pi\eta} = \begin{cases} \theta & \text{if } \eta = c_1 \cdots c_i \cdots c_k(n), \\ 1 & \text{if } \eta = c_1 \cdots c_i^j \cdots c_k, \\ 0 & \text{otherwise,} \end{cases} \tag{2.1b}$$

where

$$c_i = (n_1 \cdots n_{l(i)}) \quad \text{and} \quad c_i^j = (n_1 \cdots n_{j-1} n n_j \cdots n_{l(i)}). \tag{2.1c}$$

Finally,

$$q_{\pi\pi} = -(n-1+\theta). \tag{2.1d}$$

Now define $I(t)$ for $t \geqslant 0$ by

$$I(t) = \text{number of integers} \geqslant 1 \text{ in } \Pi(t).$$

Then $\{I(t), t \geqslant 0\}$ is precisely the linear birth process with immigration alluded to in the introduction. Its structure is determined by a sequence of independent random variables $\{\rho_n, n = 0, 1, \ldots\}$; ρ_n is the waiting time of $\{I(t)\}$ in state n, and it follows from (2.1d) that ρ_n has density

$$d_n \, \mathrm{e}^{-d_n t} \quad \text{if } t > 0, \qquad d_n = n + \theta. \tag{2.2}$$

$I(t) > n$ if, and only if, $\rho_0 + \cdots + \rho_n \leqslant t$, and the distribution of $I(t)$ is given by (1.1).

As might be expected, $\{I(t), t \geqslant 0\}$ plays the role of the time-scale of $\{\Pi(t), t \geqslant 0\}$; a new integer is added to the permutation at the times at which $I(t)$ changes state. Furthermore, standard theory applied to (2.1) shows that the jump-chain of $\{\Pi(t), t \geqslant 0\}$ is a Markov chain $\{\Pi_n, n = 0, 1, 2, \ldots\}$ on Y with $\Pi_n \in Y_n$, $\Pi_0 = (0)$ and one-step transition probabilities given by

$$P(\Pi_1 = (1) \,|\, \Pi_0 = (0)) = 1; \tag{2.3a}$$

whereas if $n > 1$, then

$$P(\Pi_n = c_1 \cdots c_i^j \cdots c_k \,|\, \Pi_{n-1} = c_1 \cdots c_i \cdots c_k) = \frac{1}{n-1+\theta}, \tag{2.3b}$$

and

$$P(\Pi_n = c_1 \cdots c_k(n) \,|\, \Pi_{n-1} = c_1 \cdots c_k) = \frac{\theta}{n-1+\theta}, \tag{2.3c}$$

the conventions in (2.1c) applying.

It will be convenient to let $|\pi|$ denote the number of cycles in the permutation π, and to let $x_{(n)} = x(x+1) \cdots (x+n-1)$. The complete stochastic structure of the process is contained in:

Theorem. (a) *$P(\Pi_n = \pi) = \theta^k / \theta_{(n)}$ if $\pi \in Y_n$ and $|\pi| = k$;*

(b) *$\{\Pi_n, n = 0, 1, \ldots\}$ and $\{I(t), t \geqslant 0\}$ are independent processes;*

(c) *$P(\Pi(t) = \pi) = (\theta^k / n!) \, \mathrm{e}^{-\theta t}(1-\mathrm{e}^{-t})^n$ if $\pi \in Y_n$ and $|\pi| = k$.*

Proof. (a) Notice first that from a given $\pi \in Y_n$ it is possible to reconstruct precisely the cyclic representation of the population at each of its previous sizes by successively removing the integers $n, n-1, \ldots, 1$ from π. Let n_j be the smallest integer in the jth cycle of π. Recall that $n_1 = 1$. Individuals $1, n_2, \ldots, n_k$ are the "founding fathers" of the k families in π; these are the only immigrants. Given these, the path to π is uniquely determined. Hence it follows from (2.3a) through (2.3c) that

$$\begin{aligned} P(\Pi_n=\pi) &= \frac{\theta}{\theta}\,\frac{1}{\theta+1}\,\frac{1}{\theta+2}\cdots\frac{1}{\theta+n_2-2}\,\frac{\theta}{\theta+n_2-1} \\ &\quad \times\frac{1}{\theta+n_2}\,\frac{1}{\theta+n_2+1}\cdots\frac{\theta}{\theta+n_3-1} \\ &\quad \times\cdots\times \\ &\quad \times\frac{1}{\theta+n_k}\cdots\frac{1}{\theta+n-1} \\ &= \frac{\theta^k}{\theta_{(n)}}. \end{aligned}$$

To prove (b), we use an argument analogous to that of Kingman (1982). Conditional on the jump chain the sojourn times are independent, the sojourn time in state π having an exponential distribution with parameter $-q_{\pi\pi}$ determined by (2.1d). If $\Pi_n=\pi$ then $-q_{\pi\pi}=d_n$ and so the conditional distribution of ρ_n, given $\Pi_n=\pi$ does not depend on π; its conditional distribution is the same as its unconditional distribution (2.2). It follows that the joint conditional distributions of $\{I(t), t\geq 0\}$ given $\{\Pi_n, n=0,1,\ldots\}$ are the same as its unconditional distributions. $\{I(t), t\geq 0\}$ and $\{\Pi_n, n=0,1,\ldots\}$ are therefore independent.

To establish (c), note that under the given conditions

$$P(\Pi(t)=\pi)=P(\Pi_n=\pi)\cdot P(I(t)=n),$$

so the result follows from (a), (b) and (1.1).

3. Some special cases

A number of results in the literature may be obtained as special cases of the permutation process. For example, we have the following proposition.

Proposition

$$P(\Pi_n \textit{ has } k \textit{ cycles, the first of size } n_1,\ldots, \textit{ the kth of size } n_k)$$

$$=\frac{\theta^k}{\theta_{(n)}}\,\frac{n!}{n_k(n_k+n_{k-1})\cdots(n_k+n_{k-1}+\cdots+n_1)}. \tag{3.1}$$

Proof. To get the required probability, we need to sum $P(\Pi_n=\pi)$ over all $\pi\in\Gamma$, the set of permutations in Y_n with k cycles, and ordered cycle sizes $n_1,\ldots,n_k$.

Since $P(\Pi_n = \pi)$ depends on π only through $|\pi|$, the required probability is

$$= \frac{\theta^k}{\theta_{(n)}} \times \text{the number of permutations in } \Gamma$$

$$= \frac{\theta^k}{\theta_{(n)}} \times \frac{(n-1)!}{(n-n_1)(n-n_1-n_2)\cdots(n-n_1-\cdots-n_{k-1})}$$

$$= \frac{\theta^k}{\theta_{(n)}} \times \frac{n!}{n_k(n_k+n_{k-1})\cdots(n_k+n_{k-1}+\cdots+n_1)}$$

and we are done.

We let $\{A(t), t \geq 0\}$ denote the process of age-ordered family sizes obtained by collapsing $\{\Pi(t), t \geq 0\}$ by ignoring which particular individuals are in which cycles. The distribution of $A(t)$ follows immediately from (3.1), (1.1) and the theorem as

$$P(A(t) = (n_1, n_2, \ldots, n_k)) = \frac{e^{-\theta t}(1-e^{-t})^n \theta^k}{n_k(n_k+n_{k-1})\cdots(n_k+\cdots+n_1)}, \tag{3.2}$$

if $n_1 + \cdots + n_k = n$.

The result of (3.2) appears in Tavaré (1987); the asymptotic behavior of the family sizes is there analysed by point-process methods. The probability distribution in (3.1) appeared in Donnelly and Tavaré (1986) in the context of population genetics models; it is the probability that a sample of n genes from a stationary infinitely-many neutral alleles model (with mutation parameter θ) has k alleles, n_1 of the oldest allele, $\ldots, n_k$ of the youngest allele. This distribution is closely related to the Ewens Sampling Formula (Ewens, 1972), which would arise in the present context as the distribution of the numbers in the families when the age-ordering of the families is ignored. The jump-chain $\{\Pi_n, n = 0, 1, \ldots\}$, or versions of it, appears in other guises; for example, it is a detailed description of the so-called 'Chinese restaurant process' described by Aldous (1985). See also Hoppe (1984), Watterson (1984), and Donnelly (1986) for related material.

The distribution of $|\Pi_n|$ immediately also follows:

$$P(|\Pi_n| = k) = \sum_{\text{all } \pi \text{ with } |\pi| = k} P(\Pi_n = \pi)$$

$$= \frac{\theta^k}{\theta_{(n)}} \times \text{the number of } n\text{-permutations with } k \text{ cycles}$$

$$= \frac{\theta^k}{\theta_{(n)}} |S_n^k|, \tag{3.3}$$

S_n^k being a Stirling number of the first kind.

The counting techniques in the previous examples can also be used to find the joint distribution of the number of individuals in the r oldest families, the marginal distribution of the number of individuals in the rth oldest family, and the conditional joint distributions of the family sizes given the total number of families. Of course, these results can be reinterpreted in terms of the structure of permutations. For example, if the immigration parameter is $\theta = 1$, then it is clear that Π_n is a random permutation of the first n integers (that is, each such permutation has probability $1/n!$). From (3.1), it follows that the probability that a random permutation of the first n integers has k cycles, n_1 in the first cycle, n_2 in the second, ..., n_k in the kth cycle is

$$\frac{1}{n_k(n_k+n_{k-1})\cdots(n_k+\cdots+n_1)},$$

cf. Vershik and Shmidt (1977).

It is, of course, tempting to use these methods to analyse birth processes with immigration (and non-linear rates) and birth-and-death processes with immigration. To a certain extent this can be done, but the results are extremely complicated.

Acknowledgement

We both thank Professor W.J. Ewens for his comments.

References

D.J. Aldous, Exchangeability and related topics, Lecture Notes in Mathematics 1117 (Springer-Verlag, New York, 1985) pp. 1-198.

P. Donnelly, Partition structures, Pólya urns, the Ewens sampling formula and the ages of alleles, Theor. Popn. Biol. 30 (1986) 271-288.

P. Donnelly and S. Tavaré, The ages of alleles and a coalescent, Adv. Appl. Prob. 18 (1986) 1-19.

W.J. Ewens, The sampling theory of selectively neutral alleles, Theor. Popn. Biol. 3 (1972) 87-112.

F. Hoppe, Pólya-like urns and the Ewens sampling formula, J. Math. Biol. 20 (1984) 91-94.

D.G. Kendall, Stochastic processes and population growth, J. Royal Statist. Soc. 11 (1949) 230-264.

J.F.C. Kingman, The coalescent, Stoch. Proc. Appln. 13 (1982) 235-248.

S. Tavaré, The birth process with immigration, and the genealogical structure of large populations, J. Math. Biol. (1987), in press.

A.M. Vershik and A.A. Shmidt, Limit measures arising in the asymptotic theory of symmetric groups I, Theory Prob. Applns., 22 (1977) 70-85.

G.A. Watterson, Estimating the divergence time of two species, Statistics Research Report, No. 94, Monash University, Australia (1984).