J. Math. Biol. (1987) 25: 161-168

Journal of
Mathematical
Biology

The birth process with immigration, and the genealogical structure of large populations

Simon Tavaré

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

Abstract. This paper studies a version of the birth and immigration process in which families are followed in the order of their appearance. This age structure is related to a number of results from population genetics, in particular the genealogical structure of the infinitely-many neutral alleles model. The asymptotic behavior of this genealogy is an easy consequence of the structure of the age-ordered family size process.

Key words: Population genetics — Genealogy — Birth process — Point process — Poisson-Dirichlet distribution

1. Introduction

This paper studies several aspects of behavior of the sizes of families that arise in the linear birth process with immigration. We focus on a stochastic process $\{A(t), t \geqslant 0\}$ on $\mathbb{R}^{\infty}$ that records the sizes of the families in the order of their appearance in the population. The marginal distribution of $A(t)$ is given explicitly in Sect. 2.

In Sect. 3 the structure of the process is decomposed into its jump chain $\{J_n, n \geqslant 0\}$ and a time-scale $\{I(t), t \geqslant 0\}$. $I(t)$ is the number of individuals in the population at time t, and J_n records the decomposition of the extant individuals into (age-ordered) family sizes. $\{J_n, n \geqslant 0\}$ has the structure of a Pólya-like urn model, and it is intimately related to the genealogical behavior of the infinitely-many neutral alleles model in population genetics theory.

In Sect. 4, the asymptotic behavior of $e^{-t}A(t)$ as $t \to \infty$ is found, and the limiting vector is analyzed by elementary point process methods. The asymptotic proportions of individuals in the population that are from the oldest, second oldest,... families have a joint distribution that is the same as a size-biased permutation of a Poisson-Dirichlet distribution. Thus the process $\{A(t), t \geqslant 0\}$ and its associated limit theory provides an elementary way to study the behavior of age-ordered samples taken from infinitely-many neutral allele models.

2. The age-ordered FS process

2.1. Preliminary properties

We record here some standard results that will be needed in the sequel. The linear birth process is a time-homogeneous Markov process taking values in the set $\mathbb{N} = \{1, 2, 3, \ldots\}$, and whose behavior is specified by the non-negative parameter λ, the birth-rate per head per unit time. The (conservative) Q-matrix of the process has elements determined by

$$q_{i,i+1} = i\lambda, \qquad i = 1, 2, \ldots,$$

the other off-diagonal elements of Q being zero. If we let $N(t)$ denote the number of individuals in the population at time t, then it is well known that

$$\mathbb{P}(N(t) = n \mid N(0) = 1) = e^{-\lambda t}(1 - e^{-\lambda t})^{n-1}, \qquad n = 1, 2, \ldots \tag{2.1}$$

cf. Kendall (1949). Now imagine that at the points $0 < T_1 < T_2 < \cdots$ of a homogeneous Poisson process of rate θ, we initiate families of individuals, each starting from a single individual and evolving (independently of each other) as a linear birth process. Let $I(t)$ be the number of individuals in the population at time t, with $I(0) = 0$. The process $\{I(t), t \geq 0\}$ is known as the linear birth process with immigration (BI). It will be convenient in what follows to scale time so that $\lambda = 1$. The distribution of $I(t)$ is then given by

$$\mathbb{P}(I(t) = n) = \binom{\theta + n - 1}{n} e^{-\theta t}(1 - e^{-t})^n, \qquad n = 0, 1, \ldots. \tag{2.2}$$

The BI process $I(\cdot)$ described above does not provide any detailed information about the growth of the families themselves. Here we introduce a process which follows the sizes of the families in the BI *in the order of their appearance in the population.* To this end, define

$$\zeta_i(t) = \begin{cases} \text{the size at time } t \text{ of the family initiated at time} \\ T_i, \quad \text{if } t \geq T_i; \\ 0, \quad \text{if } t < T_i. \end{cases}$$

The age-ordered family-size (FS) process of interest is

$$A(t) \equiv (\zeta_1(t), \zeta_2(t), \ldots), \qquad t > 0 \tag{2.3a}$$

and

$$A(0) = 0 \equiv (0, 0, \ldots) \tag{2.3b}$$

The state-space of $A(\cdot)$ is the subset $\mathcal{S}$ of non-negative integer valued sequences with all but finitely many zeros given by

$$\mathcal{S} = \{(\eta_1, \eta_2, \ldots, \eta_l, 0, 0, \ldots) : \eta_i \in \mathbb{N}, 1 \leq i \leq l, l = 1, 2, \ldots\} \cup \{0\}$$

$\{A(t), t \geq 0\}$ is a Markov process on $\mathcal{S}$, and its transition rates $\{q_{zw}\}$ are determined as follows. Suppose $z = (\eta_1, \eta_2, \ldots, \eta_l, 0, 0, \ldots) \in \mathcal{S}$. Then a transition can occur to a state

$$w = (\eta_1, \ldots, \eta_{i-1}, \eta_i + 1, \eta_{i+1}, \ldots, \eta_l, 0, 0, \ldots) \quad \text{at rate } q_{zw} = \eta_i, \tag{2.4a}$$

$i=1, 2, \ldots, l$. Alternatively, a new family can be founded, in which case a transition can occur to the state

$$w=(\eta_1, \eta_2, \ldots, \eta_l, 1, 0, 0, \ldots) \quad \text{at rate } \theta. \tag{2.4b}$$

Finally, the diagonal elements are determined from (2.4a) and (2.4b) as

$$q_{zz}=-(\theta+\eta_1+\eta_2+\cdots+\eta_l).$$

The distribution of $A(t)$ is provided by the following result.

Theorem 1. *With time scaled so that* $\lambda=1$,

$$\mathbb{P}(A(t)=0)=e^{-\theta t},$$
$$\mathbb{P}(A(t)=(\eta_1, \ldots, \eta_l, 0, 0, \ldots))=\frac{e^{-\theta t}(1-e^{-t})^{\Sigma\eta_r}\theta^l}{\eta_l(\eta_l+\eta_{l-1})\cdots(\eta_l+\cdots+\eta_1)}. \tag{2.5}$$

Remarks. Analogous results in the context of the birth and death process with immigration may be found in Theorem 2 of Tavaré (1987); Theorem 1 may either be viewed as the special case in which the death-rate is $\mu=0$, or it can be established by direct methods by the usual idea of conditioning on the number of immigrations in time t; cf. Karlin and Taylor (1981), Chap. 16.

In this paper we exploit the simpler structure of the birth and immigration process to derive (primarily) asymptotic results about the family sizes. Some of these results (in particular (3.2) and Theorem 6) are quoted without any details in Sect. 5 of Tavaré (loc. cit.).

3. The jump-chain of the age-ordered FS process

The result of Theorem 1 provides a direct construction of the distribution of the age-ordered FS process $A(\cdot)$. In this section we study in more detail the jump-chain of this process. Define $\tau_0=0$, and let τ_n be the time of the nth change of state of $A(\cdot)$. The jump-chain is the non-homogeneous Markov chain $\{J_n, n=0, 1, 2, \ldots\}$ on $\mathcal{S}$ defined by

$$J_0=0,$$
$$J_n=A(\tau_n+), \qquad n=1, 2, \ldots.$$

The state space of J_n is the set

$$\mathcal{S}_n=\left\{(\eta_1, \eta_2, \ldots)\in\mathcal{S}: \sum_{r\geq 1}\eta_r=n\right\},$$

for $n=0, 1, 2, \ldots$. From the standard theory of jump-chains, it follows that the non-zero transition probabilities of this chain are given by

$$\mathbb{P}(J_n=(\eta_1, \ldots, \eta_{i-1}, \eta_i+1, \eta_{i+1}, \ldots, \eta_l, 0, 0, \ldots) \mid J_{n-1}=(\eta_1, \eta_2, \ldots, \eta_l, 0, 0, \ldots))=\eta_i/(n-1+\theta), \quad i=1, \ldots, l; \tag{3.1}$$
$$\mathbb{P}(J_n=(\eta_1, \eta_2, \ldots, \eta_l, 1, 0, 0, \ldots) \mid J_{n-1}=(\eta_1, \eta_2, \ldots, \eta_l, 0, 0, \ldots))=\theta/(n-1+\theta),$$

for $n=1, 2, \ldots$, and $(\eta_1, \eta_2, \ldots, \eta_l, 0, 0, \ldots)\in\mathcal{S}_{n-1}$.

Theorem 2. *Let $\{J_n, n \geqslant 0\}$ be the jump-chain of $A(\cdot)$, and let $\{I(t), t \geqslant 0\}$ be the corresponding* BI *process, with $I(0)=0$. Then $I(\cdot)$ and J are independent processes, and $A(t)=J_{I(t)}$, $t \geqslant 0$.*

Furthermore,

$$\mathbb{P}(J_n=(\eta_1,\ldots,\eta_l,0,0,\ldots))=\binom{\theta+n-1}{n}^{-1}\frac{\theta^l}{\eta_l(\eta_l+\eta_{l-1})\cdots(\eta_l+\cdots+\eta_1)} \tag{3.2}$$

if $(\eta_1,\ldots,\eta_l,0,0,\ldots)\in\mathcal{S}_n$.

Proof. That $I(\cdot)$ and J are independent follows by an argument analogous to that of Kingman (1982, p. 237). Standard jump-chain theory shows that $A(t)=J_{I(t)}$. The independence result gives immediately

$$\mathbb{P}(A(t)=z)=\sum_{n\geqslant 0}\mathbb{P}(J_n=z)\mathbb{P}(I(t)=n).$$

If $z\in\mathcal{S}_n$, then $\mathbb{P}(J_r=z)=0$ unless $r=n$, in which case

$$\mathbb{P}(A(t)=z)=\mathbb{P}(J_n=z)\mathbb{P}(I(t)=n);$$

the result of (3.2) then follows directly from (2.2) and (2.5).

Remark. The distribution in (3.2) was derived (in a different context) by Donnelly and Tavaré (1986). They showed that it arises as the distribution of allelic types in an age-ordered sample from a stationary infinitely-many neutral alleles Moran model of constant size, and also in the limit of large population size in a wide variety of other genetic models.

Hoppe (1984) and Watterson (1984) describe a Pólya-like urn with a transition mechanism similar to that of $\{J_n, n \geqslant 0\}$. They focus on the fact that their models give rise to the Ewens Sampling Formula (Ewens 1972). Connections between their process and results on age-ordering (in the population genetic setting) are developed and exploited further by Donnelly (1986). See also Tavaré (1987).

Embedding the Markov chain $\{J_n, n \geqslant 0\}$ in a continuous-time Markov process $\{A(t), t \geqslant 0\}$ which has a simple structure provides an elementary way to uncover its asymptotic properties. We pursue this further in the next section.

4. The asymptotic size of families in the age-ordered FS process

We will be interested in the asymptotic size of (suitably normalized) families in the process $A(t)$ as $t \to \infty$. We recall first a result about the asymptotic behavior of linear birth processes. Let $\{N(t), t \geqslant 0;\ N(0)=1\}$ be a linear birth process of rate $\lambda=1$, as described in Sect. 2.1; we may assume that it has right-continuous sample paths. Then there exists a random variable E, having an exponential distribution with mean 1, such that

$$e^{-t}N(t)\to E \text{ almost surely,} \quad \text{as } t\to\infty. \tag{4.1}$$

(This is a special case of a standard result about the Markov branching process; cf. Athreya and Ney (1972), p. 111.)

Now let $N_1(\cdot)$, $N_2(\cdot), \ldots$ be independent copies of the linear birth process $N(\cdot)$, let E_1, $E_2, \ldots$ be independent exponential random variables satisfying $e^{-t}N_i(t) \to E_i$, a.s. as $t \to \infty$. Let T_1, $T_2, \ldots$ be the times at which families are initiated in the age-ordered FS process; $\{N_i(\cdot), i \geq 1\}$ and $\{T_i, i \geq 1\}$ are independent. Finally, let $1\{\cdot\}$ denote the indicator function of the event in $\{\ \}$.

Consider first the size $\zeta_1(t)$ of the oldest family alive at time t. Then

$$\begin{aligned} e^{-t}\zeta_1(t) &= e^{-t}N_1(t-T_1)1\{T_1 \leq t\} \\ &= e^{-T_1} e^{-(t-T_1)} N_1(t-T_1)1\{T_1 \leq t\} \\ &\to e^{-T_1}E_1, \quad \text{almost surely as } t \to \infty, \end{aligned}$$

this last following from the result of (4.1).

Theorem 3. *The age-ordered family sizes $\{\zeta_i(t)\}$ have asymptotic structure provided by*

$$e^{-t}(\zeta_1(t), \zeta_2(t), \ldots) \to (e^{-T_1}E_1, e^{-T_2}E_2, \ldots) \tag{4.2}$$

almost surely as $t \to \infty$.

Proof. Choose and fix $r \in \mathbb{N}$. Then

$$\begin{aligned} &e^{-t}(\zeta_1(t), \ldots, \zeta_r(t)) \\ &= \sum_{j=1}^{r-1} e^{-t}(N_1(t-T_1), \ldots, N_j(t-T_j), 0, 0, \ldots, 0)1\{T_j \leq t < T_{j+1}\} \\ &\quad + e^{-t}(N_1(t-T_1), \ldots, N_r(t-T_r))1\{T_r \leq t\} \\ &\to (e^{-T_1}E_1, \ldots, e^{-T_r}E_r), \quad \text{a.s. as } t \to \infty, \end{aligned} \tag{4.3}$$

by an argument analogous to that described above. Intersecting the sets on which the a.s. convergence occurs for each r provides a set of probability one on which the convergence in (4.3) holds for each $r \in \mathbb{N}$, and the theorem is proved.

The next task is to establish a useful representation for the limit random vector in (4.3). Notice first that the collection

$$\Pi = \{(T_i, E_i), i = 1, 2, \ldots\}$$

is a marked Poisson process (cf. Taylor and Karlin (1984), p. 205). That is, Π is a two-dimensional Poisson point process on $[0, \infty) \times (0, \infty)$ such that for any Borel set $S \subset [0, \infty) \times (0, \infty)$, the number of points of Π that fall in S has a Poisson distribution with mean

$$\int_S \int \theta\, e^{-y}\, dt\, dy \tag{4.4}$$

and the numbers of points falling in disjoint intervals are independent. If we take $S = \{(t, y)\colon e^{-t}y > u\}$ for any $u > 0$, then it follows from (4.4) that the number of points $(T_i, E_i) \in \Pi$ which satisfy $\exp(-T_i)E_i > u$ has a Poisson distribution with mean

$$\int_0^\infty \int_{ue^t}^\infty \theta\, e^{-y}\, dy\, dt = \int_u^\infty \frac{\theta\, e^{-v}}{v}\, dv. \tag{4.5}$$

Thus if $0<a<b$, the number of points $(T_i, E_i)\in\Pi$ for which $a<\exp(-T_i)E_i\leq b$ has a Poisson distribution with mean

$$\int_a^b \frac{\theta\, e^{-v}}{v}\, dv$$

and the numbers of such points that fall in disjoint intervals in $(0,\infty)$ are independent random variables. We have therefore proved the following result.

Theorem 4. *Let $\sigma_i = e^{-T_i}E_i$, $i=1,2,\ldots$ be the limiting random variables in Theorem 3. Then $\{\sigma_i\}$ may be viewed as the points (in some order) of a non-homogeneous Poisson process on $(0,\infty)$ with mean measure density $\theta\, e^{-x}/x$, $x>0$.*

We record one immediate consequence of the previous theorem.

Corollary 1. *The r.v. $\sigma=\sum_{i\geq 1}\sigma_i$ is a.s. finite and has a gamma density $f(x)$ given by*

$$f(x)=x^{\theta-1}\, e^{-x}/\Gamma(\theta), \qquad x>0. \tag{4.6}$$

Proof.

$$\begin{aligned}\mathbb{E}(\sigma) &= \sum_{i\geq 1}\mathbb{E}(e^{-T_i}E_i)=\sum \mathbb{E}(e^{-T_i})\mathbb{E}(E_i)\\ &= \sum_{i\geq 1}\left(\frac{\theta}{1+\theta}\right)^i\cdot 1=\theta,\end{aligned}$$

so σ is almost surely finite. And by standard manipulations with the Poisson process

$$\begin{aligned}\mathbb{E}(e^{-s\sigma}) &= \exp\left(-\theta\int_0^\infty (1-e^{-sv})\, e^{-v}v^{-1}\, dv\right)\\ &= (1+s)^{-\theta},\end{aligned}$$

which establishes the last claim; cf. Kingman (1977).

We will be interested in the asymptotic behavior of the fraction of the population that belong to the oldest, next oldest, ... families in the process. We need

Theorem 5. *Let $I(t)\equiv\sum_{j\geq 1}\zeta_j(t)$ be the total population size at time t. Then*

$$e^{-t}I(t)=\sum_{j\geq 1}e^{-t}\zeta_j(t)\to\sum_{j\geq 1}\sigma_j=\sigma, \quad \textit{a.s. as } t\to\infty.$$

Proof. First, the process $\{e^{-t}I(t), t\geq 0\}$ is a (non-negative) submartingale with respect to its natural history, and it has uniformly bounded means ($\leq\theta$ in this case, because $I(0)=0$). Hence there is a random variable I, say, such that $e^{-t}I(t)\to I$ a.s. as $t\to\infty$. From (2.2), and a simple Laplace transform argument, it follows that I has a gamma distribution with density (4.6); it has mean θ. By Fatou's Lemma, $I\geq\sigma$. But by Corollary 1, σ also has a gamma distribution with density (4.6). Hence $0\leq\mathbb{E}(I-\sigma)=\mathbb{E}I-\mathbb{E}\sigma=0$, and so $I=\sigma$ a.s.

Corollary 2.

$$I(t)^{-1}(\zeta_1(t), \zeta_2(t), \ldots) \to (P_1, P_2, \ldots) \quad \textit{a.s. as } t \to \infty,$$

where

$$P_i = \sigma_i/\sigma = e^{-T_i}E_i \Big/ \Big(\sum_{j \geq 1} e^{-T_j}E_j \Big), \qquad i = 1, 2, \ldots \tag{4.7}$$

Proof. Immediate from Theorems 4 and 5.

The random variable P_i is the asymptotic fraction of the population that belongs to the *i*th oldest family. The structure of $\{P_i\}$ is given by

Theorem 6. *Let* $\{Z_i, i = 1, 2, \ldots\}$ *be an i.i.d. sequence of random variables with common density* $g(x)$ *given by*

$$g(x) = \theta(1-x)^{\theta-1}, \qquad x \in (0, 1).$$

Then

$$P_i =^d (1-Z_1)(1-Z_2) \cdots (1-Z_{i-1})Z_i, \qquad i = 1, 2, \ldots. \tag{4.8}$$

Proof. For any $r \in \mathbb{N}$, the joint distribution of $(\sigma_1, \ldots, \sigma_r, \sum_{j>r} \sigma_j)$ is that of

$$\Big(e^{-T_1}E_1, \ldots, e^{-T_r}E_r, \sum_{j>r} e^{-T_j}E_j \Big) = \Big(e^{-T_1}E_1, \ldots, e^{-T_r}E_r, e^{-T_r} \sum_{j>r} e^{-(T_j - T_r)}E_j \Big)$$
$$=^d (e^{-T_1}E_1, \ldots, e^{-T_r}E_r, e^{-T_r}\sigma^*), \tag{4.9}$$

where σ^* has the same distribution as σ, and is independent of $E_1, \ldots, E_r$, $T_1, \ldots, T_r$. A calculation shows that the joint density of the random variables in (4.9) is given by

$$f(x_1, \ldots, x_{r+1}) = \frac{\theta^r x_{r+1}^\theta \exp(-x_1 - \cdots - x_{r+1})}{\Gamma(\theta)x_{r+1}(x_{r+1}+x_r) \cdots (x_{r+1} + \cdots + x_1)}.$$

Hence the joint density of $(\sigma_1/\sigma, \ldots, \sigma_r/\sigma, \sigma)$ is

$$g(x_1, \ldots, x_{r+1}) = \frac{(1-x_1 - \cdots - x_r)^\theta \, \theta^r}{(1-x_1) \cdots (1-x_1 - \cdots - x_r)} \, \frac{x_{r+1}^{\theta-1} \exp(-x_{r+1})}{\Gamma(\theta)}.$$

Thus σ is independent of $(\sigma_1/\sigma, \ldots, \sigma_r/\sigma)$, and a simple calculation shows that the left-hand term above is precisely the joint density of $(P_1, \ldots, P_r)$ defined in (4.8).

Remarks. There are many interconnections between the random variables P_i in (4.8), and the Poisson-Dirichlet distribution which plays so important a role in the theory of neutral mutation in population genetics; see Kingman (1977) for example. Kingman (1975) established that if $\sigma_{(1)} \geq \sigma_{(2)} \geq \cdots \geq 0$ are the ordered points of a Poisson process of the type described in Theorem 4, then $(\sigma_{(1)}/\sigma, \sigma_{(2)}/\sigma, \ldots)$ has the Poisson-Dirichlet distribution with parameter θ. Patil and Taillie (1977) showed that the size-biased permutation of a Poisson-Dirichlet distribution has the representation $(P_1, P_2, \ldots)$ given in (4.8). The interconnections between size-biasing, and the age-ordered frequencies in the infinitely-many neutral alleles model are described in Donnelly (1986).

The stochastic structure of the urn model $\{J_n, n \geqslant 0\}$ arises directly from a consideration of the genealogy of age-ordered samples taken from a stationary infinitely-many neutral alleles model; see Donnelly and Tavaré (1986). We may therefore use the results of this section to analyze the asymptotic behavior of this genealogy. Here is an example.

Corollary 3. *$n^{-1}J_n \to (P_1, P_2, \ldots)$ a.s. as $n \to \infty$, the random variables $\{P_n\}$ being given in* (4.8).

Proof. This follows immediately from Corollary 2, since if τ_n is the time of the nth change of state of $A(\cdot)$, then $J_n = A(\tau_n +)$, $I(\tau_n +) = n$, and $\tau_n \to \infty$ a.s. as $n \to \infty$.

Donnelly and Tavaré (1986) established that the asymptotic (as $n \to \infty$) fractions of a sample of size n from a stationary infinitely-many neutral alleles model that may be assigned to the oldest allelic type, the next oldest allelic type, and so on has the representation (4.8). The representation developed here may be viewed as a simple alternative method to study such distributions.

Acknowledgements. This research was supported in part by N.S.F. grant DMS 86-08857.

References

1. Athreya K. B., Ney P. E.: Branching processes. Berlin Heidelberg New York: Springer 1972
2. Donnelly P. J.: Partition structures, Pólya urns, the Ewens sampling formula and the ages of alleles. Theor. Popul. Biol. **30**, 271-288
3. Donnelly P. J., Tavaré S.: The ages of alleles and a coalescent. Adv. Appl. Prob. **18**, 1-19 (1986)
4. Ewens W. J.: The sampling theory of selectively neutral alleles. Theor. Popul. Biol. **3**, 87-112 (1972)
5. Hoppe F. M.: Pólya-like urns and the Ewens sampling formula. J. Math. Biol. **20**, 91-94 (1984)
6. Karlin S., Taylor H. M.: A second course in stochastic processes. New York: Academic Press 1981
7. Kendall D. G.: Stochastic processes and population growth. J. Roy. Statist. Soc. B **11**, 230-264 (1949)
8. Kingman J. F. C.: Random discrete distributions. J. Roy. Statist. Soc. B **37**, 1-22 (1975)
9. Kingman J. F. C.: The population structure associated with the Ewens sampling formula. Theor. Popul. Biol. **11**, 274-283 (1977)
10. Kingman J. F. C.: The coalescent. Stoch. Proc. Appln. **13**, 235-248 (1982)
11. Patil G. P., Taillie C.: Diversity as a concept and its implications for random communities. Bull. Internat. Stat. Inst. **47**, 497-515 (1977)
12. Tavaré S.: The genealogy of the birth, death and immigration process. In: Feldman M. W. (ed.) Mathematical evolutionary theory. Princeton University Press, in press (1987)
13. Taylor H. M., Karlin S.: An introduction to stochastic modeling. Orlando: Academic Press 1984
14. Watterson G. A.: Estimating the divergence time of two species. Statistics Research Report, No. 94, Monash University, Australia (1984)

Received September 22, 1986/Revised January 5, 1987