J. Math. Biol. (1987) 25: 123–159

Journal of
Mathematical
Biology

The sampling theory of neutral alleles and an urn model in population genetics*

Fred M. Hoppe

Departments of Mathematics and Statistics, The University of Michigan,
Ann Arbor, MI 48109-1027, USA

Abstract. The behaviour of a Pólya-like urn which generates Ewens' sampling formula in population genetics is investigated. Connections are made with work of Watterson and Kingman and to the Poisson-Dirichlet distribution. The order in which novel types occur in the urn is shown to parallel the age distribution of the infinitely many alleles diffusion model and consequences of this property are explored. Finally the urn process is related to Kingman's coalescent with mutation to provide a rigorous basis for this parallel.

Key words: Ages of alleles — Coalescent — Ewens' sampling formula — Genealogy — Mutation — Poisson-Dirichlet — Pólya urn — Population genetics — Size-biased sampling

1. Introduction

The following Pólya-like urn process $\mathscr{U}$ was described in Hoppe (1984). An urn contains one black ball whose mass is $\theta > 0$ and various numbers of other balls having assorted colours (non-black) each of mass one. At each instant of discrete time a ball is drawn at random (that is in proportion to its mass). If the selected ball is black then it is returned together with one additional ball of a previously unused colour. Otherwise it is returned together with one additional ball of the same colour. All new balls have unit mass and for definiteness the natural numbers will be used sequentially to label the colours as the need arises. At the outset there is a single black ball and no others in the urn. The random variable X_n is defined to be the label of the additional ball returned after the nth drawing. K is the random number of distinct labels present after the nth return and $S_i(n)$, $1 \leq i \leq K$, is the number of balls labelled i. Also $a = (a_1, a_2, \ldots, a_n)$ denotes, in the terminology of genetics, the allelic partition (Kingman (1978)) a_j being the number of times the integer j appears in the set $\{S_1(n), \ldots, S_K(n)\}$.

* This research was partially supported by the Sloan Foundation under Grant 85-6-14 and by the National Science Foundation

The use of allelic partitions is convenient in those genetic applications where no biological significance is to be attached to the actual labels which merely represent distinct allelic forms of a gene and a_j becomes the number of alleles having j copies in a sample. The sequence $\{X_1, \ldots, X_n\}$ determines a random allelic partition Π_n pertaining to which the following was established.

Theorem A. *$\{\Pi_n\}$ is a Markov process having marginal distributions*

$$\Pr[\Pi_n = a] = \frac{n!}{[\theta]^n} \prod_{i=1}^{n} \frac{\theta^{a_i}}{i^{a_i} a_i!} \tag{1.1}$$

where $[\theta]^n = \theta(\theta+1) \cdots (\theta+n-1)$ is the ascending factorial and $a = (a_1, \ldots, a_n)$ with $\sum ia_i = n$.

The right-hand side of (1.1), known as Ewens' sampling formula, was derived (Ewens (1972), see also Karlin and McGregor (1972)) as the limiting distribution of allelic numbers in a random sample, first for a discrete Wright-Fisher model with mutation and then later by various authors for other models displaying infinitely many selectively neutral alleles. The goal of this paper is to understand why (1.1) occurs in the present context where there appears neither a genetic nor a population structure.

The paper is a mixture of theorems and discussion and it is organized as follows. In Sect. 2 we present our first two theorems describing the limiting behaviour of $\mathcal{U}$. The distributions which emerge relate intimately to work of Watterson and Kingman, connections which are explored in Sects. 3 and 4 dealing with size-biased relabelling of Pólya urns and the Poisson-Dirichlet distribution, respectively. Section 4 also addresses modelling issues associated with samples from a population. In Sect. 5 we draw upon a remarkable equivalence between the order that alleles are sampled and their age distribution in order to derive various age properties in an entirely new and simple fashion. In Sect. 6 we develop a representation for Ewens' partition as a residual allocation model. Section 7 describes affinities between the Markovian property of Π_n and partition structures, while Sect. 8 establishes the relationship between the urn model in reverse time and the genealogy of the coalescent (Kingman (1982), Tavaré (1984), Watterson (1984)) with mutation. The final section summarizes the main results.

2. Limit behaviour of the urn

Theorem 1.

$$\lim_{n\to\infty} \frac{S_i(n)}{n} = P_i \quad \textit{and} \quad \sum_{i=1}^{\infty} P_i = 1 \quad \textit{a.s.}$$

Theorem 2. *The proportions*

$$Z_n = P_n \Big/ \sum_{i=n}^{\infty} P_i, \qquad n \geq 1$$

are independent and identically distributed each with a Beta $(1, \theta)$ *density* $\theta(1-z)^{\theta-1}1(0<z<1)$.

Corollary.

$$P_1 = Z_1 \quad \textit{and} \quad P_n = Z_n \prod_{i=1}^{n-1} (1-Z_i) \textit{ for } n \geq 2. \tag{2.1}$$

To prepare for the proofs we need to establish some facts concerning multivariate Pólya urns for which we refer to Blackwell and MacQueen (1973) whose elegant treatment motivated our approach.

At time $t=0$ an urn contains α_i balls of type $i(1 \leq i \leq K)$. If α_i is non-integer we interpret the fractional part as being represented by a ball having the fractional part as mass. At each instant of discrete time a ball is drawn at random (with probability proportional to its mass). The ball is then returned to the urn together with one additional ball of the same type and unit mass. The random variable X_n is the type of the nth additional ball placed into the urn. It is possible to formulate this process without using physics concepts such as mass and we may equivalently consider a fixed array of K cells labelled $1, 2, \ldots, K$ and associated non-negative reals α_i, $1 \leq i \leq K$, with $\sum \alpha_i = \theta$. Tokens are sequentially thrown at this array in such a way that if X_n denotes the cell into which the nth token lands then

$$\Pr[X_{n+1} = i \mid X_1, X_2, \ldots, X_n] = \frac{\alpha_i + \nu_i(n)}{\theta + n} \tag{2.2}$$

where $\nu_i(n)$ equals the number of tokens in cell i after n throws. The process $\{X_1, X_2, \ldots\}$ is endowed with the following properties:

$$\lim_{n\to\infty} \frac{\nu_i(n)}{n} = P_i \quad \text{and} \quad \sum P_i = 1 \quad \text{a.s.}; \tag{2.3}$$

$$\underset{\sim}{P} \equiv (P_1, P_2, \ldots, P_K) \tag{2.4}$$

has a Dirichlet distribution with parameter $(\alpha_1, \ldots, \alpha_K)$ which we denote by $D(\alpha_1, \ldots, \alpha_K)$;

Conditional on $\underset{\sim}{P}$, $\{X_1, X_2, \ldots\}$ are independent and identically distributed with distribution $\Pr[X_1 = i \mid \underset{\sim}{P}] = P_i$. (2.5)

The Dirichlet $D(\alpha_1, \ldots, \alpha_K)$ is defined as the joint distribution of a random vector of proportions $\underset{\sim}{P} \equiv (P_1, P_2, \ldots, P_K)$ where $P_i = X_i / \sum_{j=1}^{K} X_j$ and $X_1, X_2, \ldots, X_K$ are independent Gamma random variables having respective densities

$$\frac{1}{\Gamma(\alpha_i)} e^{-x_i} x_i^{\alpha_i - 1} 1(x_i > 0).$$

(It is convenient to regard the X_i as representing on some scale the abundances of different resources so that $\sum X_j$ is the total abundance and P represents the relative proportions.) The distribution is singular with respect to Lebesgue measure in K dimensions, but the $K-1$ dimensional distribution of $(P_1, P_2, \ldots, P_{K-1})$ is absolutely continuous with density

$$\frac{\Gamma(\sum_1^K \alpha_i)}{\prod_1^K \Gamma(\alpha_i)}\left(1-\sum_1^{K-1} p_i\right)^{\alpha_K-1}\prod_{i=1}^{K-1} p_i^{\alpha_i-1} \tag{2.6}$$

over the simplex $\{(p_1, \ldots, p_{K-1})\colon p_i \geq 0 \text{ and } \sum_1^{K-1} p_i \leq 1\}$. We record here for later recall the fact that the marginals P_i have Beta$(\alpha_i, \sum_{j\neq i} \alpha_j)$ densities

$$\frac{\Gamma(\sum \alpha_j)}{\Gamma(\alpha_i)\Gamma(\sum_{j\neq i}\alpha_j)}\, p_i^{\alpha_i-1}(1-p_i)^{(\sum_{j\neq i}\alpha_j)-1}1(0<p_i<1)$$

with $E(P_i)=\alpha_i/\sum_{j=1}^K \alpha_j$. A detailed discussion of the Dirichlet distribution may be found in Wilks (1962, Sect. 7.7). For notational convenience we denote the symmetric Dirichlet in which all $\alpha_i \equiv \alpha$ by the symbol $D(\alpha; K)$.

Proof of Theorem 1. Between selections of the black ball $\{X_n\}$ behaves like a Pólya urn, the types being represented by the currently used colours. Introduce $\{\lambda_i\}_{i=1}^{\infty}$ an independent family of Bernoulli random variables such that $\Pr[\lambda_i = 1] = \theta/(\theta+i-1)$ and $\Pr[\lambda_i = 0] = (i-1)/(\theta+i-1)$. Define stopping times $\{t_i\}$ by $t_1 = 1$ a.s. and, for $i \geq 2$, $t_i = \min\{j > t_{i-1}\colon \lambda_j = 1\}$ or ∞ if no such j exists. These times herald the introduction of new colours.

Fix an integer $k \geq 1$ and set for each $n \geq 1$

$$U_n^{(k)} = \begin{cases} j & \text{if } 1 \leq j \leq k,\ \lambda_j = 1 \text{ and } X_{n+k} = X_j \\ 0 & \text{else.} \end{cases}$$

The process $\{U_n^{(k)}\}_{n=1}^{\infty}$ takes values in $\{0, 1, 2, \ldots, k\}$ and conditional on $\{X_1, \ldots, X_k\}$ it behaves like a $(k+1)$ colour Pólya urn where the types have been labelled according to the occurrence time of their initial appearance if that time was prior to k or zero (otherwise). The initial urn composition for the $\{U_n^{(k)}\}$ process is the relabelled configuration of the $\{X_n\}$ process after k draws and is given by $(\alpha_0^{(k)}, \ldots, \alpha_k^{(k)})$ where

$$\alpha_i^{(k)} = \begin{cases} \lambda_i S_{X_i}(k) & 1 \leq i \leq k, \\ \theta & i = 0. \end{cases}$$

Also, let

$$S_i(n, k) = \#\{j\colon U_j^{(k)} = i,\ 1 \leq j \leq n\} \quad \text{if } 0 \leq i \leq k$$

denote the relabelled configuration of observed types at time $n+k$ (where all types first observed after time k have been labelled zero, as noted previously). Most of the $k+1$ colours are of course absent.

Let $n \to \infty$ and apply (2.3) to deduce the existence of random quantities $\{\gamma_i^{(k)}; 0 \leq i \leq k\}$ such that

$$\Pr\left[\lim_{n\to\infty} \frac{S_i(n, k)+\alpha_i^{(k)}}{n+k+\theta} = \gamma_i^{(k)} \,|\, X_1, \ldots, X_k\right] = 1$$

and take expectations to obtain

$$\lim_{n\to\infty} \frac{S_i(n, k)+\alpha_i^{(k)}}{n+k+\theta} = \gamma_i^{(k)} \quad \text{a.s.}$$

Observe that $S_i(n, k)+\alpha_i^{(k)}=\alpha_i^{(n+k)}$ for $1\leqslant i\leqslant k$ by definition so that

$$\lim_{n\to\infty}\frac{\alpha_i^{(n+k)}}{n+k}=\gamma_i^{(k)}$$

implying that $\gamma_i^{(k)}$ does not depend on k and giving

$$\lim_{n\to\infty}\frac{\alpha_i^{(n)}}{n}=\gamma_i \quad \text{a.s.} \qquad 1\leqslant i<\infty$$

where $\gamma_i^{(k)}\equiv\gamma_i$ for $1\leqslant i<\infty$.

Next, note that

$$E[\gamma_0^{(k)}\,|\,X_1, X_2, \ldots, X_k]=\frac{\theta}{\theta+k}$$

because conditional on $(X_1, X_2, \ldots, X_k)$ the random variable $\gamma_0^{(k)}$ has a Beta-(θ, k) distribution being a marginal of the Dirichlet distribution $D(\alpha_0^{(k)}, \alpha_1^{(k)}, \ldots, \alpha_k^{(k)})$. Thus

$$E\left[1-\sum_{i=1}^{k}\gamma_i\right]=\frac{\theta}{\theta+k}.$$

Since

$$1-\sum_{i=1}^{\infty}\gamma_i=\lim_{k\to\infty}1-\sum_{i=1}^{k}\gamma_i$$

we use dominated convergence to force

$$E\left[1-\sum_{i=1}^{\infty}\gamma_i\right]=\lim_{k\to\infty}\frac{\theta}{\theta+k}=0$$

and because $1-\sum_{i=1}^{\infty}\gamma_i\geqslant 0$ this assures $\sum_{i=1}^{\infty}\gamma_i=1$.

With the limit behaviour in hand for the relabelled process we return to the original sequence $\{X_n\}$. The colour i first appears at time t_i and thus $S_i(n)=\alpha_{t_i}^{(n)}$ for $n\geqslant t_i$. We conclude that

$$\lim_{n\to\infty}\frac{S_i(n)}{n}=\gamma_{t_i}\equiv P_i \quad \text{a.s.}$$

By our construction the only positive terms in the sequence $(\gamma_1(\omega), \gamma_2(\omega), \ldots)$ are $(\gamma_{t_1}(\omega), \gamma_{t_2}(\omega), \ldots)$ and thus $\sum_{i=1}^{\infty}P_i=\sum_{i=1}^{\infty}\gamma_{t_i}=1$ because $\Pr[t_i<\infty$ finitely often$]=\lim_{n\to\infty}\Pr[\lambda_i=0, \forall i\geqslant n]=\lim_{n\to\infty}\prod_{i=n}^{\infty}(i-1)/(\theta+i-1)=0$.

Proof of Theorem 2. If $(Y_1, \ldots, Y_{n+1})$ have a joint $D(\alpha_1, \alpha_2, \ldots, \alpha_{n+1})$ distribution then the random variables $\{U_1, \ldots, U_n\}$, where $U_k=Y_k/\sum_{i=k}^{n+1}Y_i$, are independent with Beta$(\alpha_k, \sum_{i=k+1}^{n+1}\alpha_i)$ densities respectively (Connor and Mosimann (1969)). We will use this fact to verify that Z_n and the family $\{Z_1, \ldots, Z_{n-1}\}$ are independent for all $n\geqslant 2$. Thus fixing n denote by $\mathscr{F}^{(n)}$ the σ-algebra generated by $\{X_1, X_2, \ldots\}$ up to the time t_n, let $\alpha_k=S_k(t_n)$, $1\leqslant k\leqslant n$, represent the number of balls labelled k in the urn just after the first ball with label n is added and define $\alpha_{n+1}=\theta$. By coalescing into one group all types with labels $\geqslant n+1$ we obtain a process whose probabilistic behaviour from time t_n onwards is that of a Pólya urn with initial composition $(\alpha_1, \ldots, \alpha_{n+1})$. Accordingly, given $\mathscr{F}^{(n)}$ the joint distribution of $(P_1, \ldots P_n, \sum_{i=n+1}^{\infty} P_i)$ is $D(\alpha_1, \alpha_2, \ldots, \alpha_{n+1})$ and consequently given $\mathscr{F}^{(n)}$ the random variables $\{Z_i, 1 \leq i \leq n\}$ are independent Beta(α_i, β_i) where $\beta_i = \sum_{j=i+1}^{n+1} \alpha_j$. The definition of t_n requires $\alpha_n = 1$ a.s. forcing the joint density given $\mathscr{F}^{(n)}$ to factor into a product of two terms,

$$\left[\prod_{i=1}^{n-1} \frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)} z_i^{\alpha_i - 1}(1 - z_i)^{\beta_i - 1}\right] \quad \text{and} \quad \theta(1 - z_n)^{\theta - 1}.$$

The unconditional joint density likewise factors as $\phi(z_1, \ldots, z_{n-1})\theta(1 - z_n)^{\theta - 1}$ for some function ϕ. This displays $\theta(1 - z_n)^{\theta - 1}$ as the marginal density of Z_n and shows Z_n and $\{Z_1, \ldots, Z_{n-1}\}$ to be independent. Since n is arbitrary the proof is complete.

Proof of Corollary.

$$\begin{aligned} P_n &= Z_n \sum_{i=n}^{\infty} P_i \\ &= Z_n(1 - P_1 - \cdots - P_{n-1}) \\ &= Z_n \frac{1 - P_1 - \cdots - P_{n-1}}{1 - P_1 - \cdots - P_{n-2}} (1 - P_1 - \cdots - P_{n-2}) \\ &= Z_n(1 - Z_{n-1})(1 - P_1 - \cdots - P_{n-2}). \end{aligned}$$

Induction finishes up the proof.

3. Connection with Watterson's k-allele model

The genetics underlying (1.1) is obscured by the above formulation. In the original context (Ewens (1972)) a population of $2N$ genes with an infinite number of possible alleles at a locus with no selective differences is reproducing according to a discrete time Wright–Fisher process. The $2N$ genes of each generation are formed by sampling with replacement from the gene pool of the previous generation. Additionally as each gene is selected there is a probability u that a (non-recurrent) mutation occurs to a novel allele. The process tracking the allele numbers is transient since each allele is eventually lost, but the partition of the population forms a finite state irreducible Markov chain which approaches an equilibrium distribution under mutation and genetic drift from which a sample of size n is taken whose partition distribution (in the limit $N \to \infty$, $u \to 0$, $4Nu \to \theta$) is then shown to be described by (1.1).

Strictly speaking, if a sample is taken then there must be a well-defined population (probability model), namely the equilibrium distribution. Unfortunately (Ewens (1979)) this distribution is not tractable and Karlin and McGregor (1972) bypass it to derive (1.1) instead analysing the line of descent from one generation to the next to derive a recursion which simplifies in the limit $4Nu \to \theta$.

A variety of different models for selectively neutral mutation lead, in the limit of increasing population size to (1.1). Kingman (1978) has described three broad features linking the models and Watterson (1976) and Kingman (1977) have identified the limiting population as being Poisson–Dirichlet, a type of limit of Dirichlet populations. (This will be defined below.) In this section we explain how the urn model $\mathscr{U}$ arises as a limit of samples from Dirichlet populations thereby recovering the genesis of (1.1) as resulting from a sample. We rely on an approach due to Watterson (1976) to the infinite alleles model.

Watterson begins with a k-allele Wright–Fisher process allowing a constant mutation probability from one type A_i to any other type A_j. For large population sizes he approximates by a diffusion whose stationary distribution of allele frequencies is the symmetric Dirichlet $D(\alpha; k)$ where $\alpha = \theta/(k-1)$ and θ is a parameter determined by the diffusion. Watterson shows that in the limit ($k \to \infty$, $\alpha \to 0$, $k\alpha \to \theta$) the distribution of the partition determined by a random sample $\{X_1, X_2, \ldots, X_n\}$ from $D(\alpha; k)$, converges for each fixed n, to the right-hand side of (1.1). (Here X_i takes the value $j (1 \leqslant j \leqslant k)$ if the ith observation is of allele A_j.) Now such a sequence $\{X_1, X_2, \ldots, X_n\}$ can be generated from an urn. In fact, relations (2.2)–(2.5) exhibit a duality, familiar to enthusiasts of exchangeability, identifying a random sample from a Dirichlet population as a sequence of selections from an appropriate multivariate Pólya urn. We may therefore imagine a Pólya urn as described in Sect. 2, having an initial composition of α balls for each of the k allelic types, from which n drawings are made. This sequence has the same joint distribution as a random sample from $D(\alpha; k)$ and thus has the same partition distribution.

The observations $\{X_1, X_2, \ldots, X_n\}$ are exchangeable so the information in them may be summarized by the set of occupancy numbers $\{n_1, n_2, \ldots, n_k\}$ where n_i denotes the number of observations of allele A_i. Thus we can resort to another method of bookkeeping whereby we record the observations by relabelling the alleles according to the order in which they first appear in the sample. Thus the first allele observed is assigned the new label 1 and so are all subsequently observed alleles indistinguishable from it. The second different allele observed is relabelled 2 and this process is continued to relabel all k types. Represent this relabelled sequence as $\{Y_1, Y_2, \ldots, Y_n\}$. The occupancy numbers of this sequence obviously form a permutation of $(n_1, n_2, \ldots, n_k)$ and thus the partition is the same.

The probabilistic structure of $\{Y_1, Y_2, \ldots\}$ is most transparent by considering the observations as being generated by the Pólya urn described two paragraphs up. At the start the total mass of all the balls in the urn is $k\alpha$. One ball is selected, relabelled as 1, and returned together with an additional ball also labelled 1. Moreover all of the original balls in the urn of the same type as the ball selected are renamed 1. This gives $Y_1 = 1$ a.s. The urn now contains $k\alpha + 1$ balls of which $(\alpha + 1)$ carry the new label 1 while the remainder have not been renamed. Thus with probability $(\alpha+1)/(k\alpha+1)$ Y_2 takes the value 1 while with probability $(k-1)\alpha/(k\alpha+1)\, Y_2$ will take the value 2 (corresponding to a different allele). In general, given $\{Y_1, Y_2, \ldots, Y_n\}$

$$\Pr[Y_{n+1} = i \mid Y_1, Y_2, \ldots, Y_n] = \frac{\alpha + \mu_i(n)}{k\alpha + n} \tag{3.1}$$

where $\mu_i(n) = \#\{1 \leqslant j \leqslant n \colon Y_j = i\}$ and therefore

$$\Pr[Y_{n+1} = \text{new label} \mid Y_1, Y_2, \ldots, Y_n] = \frac{(k - \phi(n))\alpha}{k\alpha + n} \tag{3.2}$$

where $\phi(n)$ is the number of distinct values among $\{Y_1, \ldots, Y_n\}$. Equations (3.1) and (3.2) describe an urn model which, for large k and small α with $k\alpha \to \theta$, closely approximates $\mathcal{U}$ and we refer to it as $\mathcal{U}(\alpha; k)$.

We interpret these equations as follows. Each time an allele type is first drawn from the urn, the mass of alleles not yet observed decreases by an amount α. Ultimately the source of new alleles is exhausted since $k < \infty$. But when k is large and α small, the effective mass of unobserved alleles remains nearly constant, for moderate sample sizes, at the nominal value $k\alpha$. Since in the limit ($k \to \infty$, $\alpha \to 0$ and $k\alpha \to \theta$), the probabilities in (3.1) and (3.2) converge, respectively, to $\mu_i(n)/(\theta+n)$ and $\theta/(\theta+n)$ precisely their counterparts defining $\mathcal{U}$, we would expect that the corresponding partition distributions converge to the partition distribution of the urn $\mathcal{U}$, which, by Watterson's (1976) result is given by (1.1). This can be made rigorous using weak convergence as for example in the proof of the theorem in Kingman (1977). However, the truth of this assertion is already known by the direct, although non-intuitive combinatorial argument in Hoppe (1984) and our goal here is merely to provide some insight into Theorem A. The urn model $\mathcal{U}$ thus represents a limiting case of sampling from a Dirichlet population, by maintaining a constant source θ of novel alleles regardless of the number of distinct alleles already present in the sample.

In a similar fashion we can provide illumination for Theorems 1 and 2. Observe that (2.3) is a strong law of large numbers recovering the Dirichlet distribution from the limiting frequencies. Equally there is a limit for the relative frequencies in the relabelled $\{Y_1, Y_2, \ldots\}$ and we proceed to show this limiting population to be the size-biased permutation of $D(\alpha; k)$. We first define this concept.

Let $\underset{\sim}{P} = (P_1, P_2, \ldots)$ be a random probability distribution on the integers,

$$P_i \geq 0 \quad \text{and} \quad \sum_i P_i = 1 \quad \text{a.s.}$$

$\underset{\sim}{P}$ may be supported (as with $D(\alpha; k)$) on a finite subset of the integers and this subset may be random. For concreteness of terminology we will continue to call P_i the frequency of some allele A_i in a hypothetical population. The size-biased permutation $\underset{\sim}{P}^s \equiv (P_1^s, P_2^s, \ldots)$ randomly rearranges the frequencies $\{P_i\}$ in proportion to their values, that is for any $i \geq 1$ and distinct subscripts $\sigma(1)$, $\sigma(2), \ldots, \sigma(i)$,

$$\Pr[P_j^s = P_{\sigma(j)}: 1 \leq j \leq i \,|\, \underset{\sim}{P}] = P_{\sigma(1)} \prod_{j=2}^{i} P_{\sigma(j)} \left(1 - \sum_{k=1}^{j-1} P_{\sigma(k)}\right)^{-1}. \tag{3.3}$$

This equation has a simple interpretation. We imagine that an individual is randomly selected from the population and we denote the frequency of its (random) allelic type for P_1^s. All such alleles are then removed from the population after which another random selection is made. The frequency (in the entire population) of this second randomly chosen allele is P_2^s and this process is repeated indefinitely or until all alleles have been exhausted.

If the population is finite, then each random selection is made using a uniform distribution over all remaining individuals. This means that an allele is selected in direct proportion to its numbers in the population. On the other hand, when the type space is countably infinite, as there is no uniform distribution over such a set, the selection procedure is interpreted as meaning directly that an allele type is chosen in proportion to its frequency. Hence the name size-biased permutation.

There is an equivalent mechanism for generating $\underset{\sim}{P}^s$ from a random sample (with replacement) $\{X_1, X_2, \ldots\}$ from $\underset{\sim}{P}$. (Here again X_i is directed to take the value j if the ith observation is allele A_j.) Denote by Q_j the relative frequency in the population of the jth distinct allele in the sample. Clearly if $\sigma(1)$ and $\sigma(2)$ are any distinct positive integers

$$\Pr[Q_1 = P_{\sigma(1)} \,|\, \underset{\sim}{P}] = P_{\sigma(1)}$$

and

$$\Pr[Q_1 = P_{\sigma(1)}, Q_2 = P_{\sigma(2)} \,|\, \underset{\sim}{P}] = \sum_{k=1}^{\infty} P^k_{\sigma(1)} P_{\sigma(2)} = \frac{P_{\sigma(1)} P_{\sigma(2)}}{1 - P_{\sigma(1)}}.$$

These are just (3.3) for $i = 1, 2$, and it can be shown more generally (by induction) that $\underset{\sim}{Q} = (Q_1, Q_2, \ldots)$ is a version of $\underset{\sim}{P}^s$.

In a correspondingly natural way, mimicking an earlier construction, the sample $\{X_1, X_2, \ldots\}$ determines a sequence $\{Y_1, Y_2, \ldots\}$ which relabels the alleles successively with the positive integers according to the order they first appear in the sample.

Let $Q_j^{(n)} = (1/n) \# \{1 \leq i \leq n \colon Y_i = j\}$ be the empirical probability function for the relabelled observations. The strong law for exchangeable variables (2.3) asserts that

$$\lim_{n \to \infty} \frac{1}{n} \# \{1 \leq i \leq n \colon X_i = k\} \quad \text{exists for each } k.$$

Hence for almost all realizations $\{X_1(\omega), X_2(\omega), \ldots\}$ the proportions of all the different types converge. The labelling is irrelevant for the convergence along each sample path. Therefore

$$\lim_{n \to \infty} Q_j^{(n)} \quad \text{exists for each } j.$$

But by definition, Q_j is the relative frequency in the population of the jth distinct allele in the sample and thus

$$\lim_{n \to \infty} Q_j^{(n)} = Q_j. \tag{3.4}$$

When $\underset{\sim}{P}$ is the symmetric Dirichlet $D(\alpha; k)$ then this method of generating the size-biased permutation $\underset{\sim}{P}^s$ is by its construction equivalent to the urn $\mathcal{U}(\alpha; k)$ and (3.4) thus describes the limiting proportions of the k types in $\mathcal{U}(\alpha; k)$. Since $\mathcal{U}(\alpha; k)$ approximates $\mathcal{U}$, we would therefore expect $\underset{\sim}{P}^s$ to be close to the population defined by (2.1). This is confirmed by the representation (Patil and Taillie (1977))

$$P_1^s = U_1, \qquad P_i^s = U_i \prod_{j=1}^{i-1} (1 - U_j), \qquad 2 \leq i \leq k-1 \tag{3.5}$$

of the size-biased permutation of the symmetric Dirichlet, where $\{U_i\}$ are independent random variables having Beta$(1 + \alpha, (k - i)\alpha)$ distributions respectively ($\underset{\sim}{P}^s$

is an example of a completely neutral vector, in the terminology of Connor and Mosimann (1969), or a residual allocation model, in the terminology of Patil and Taillie (1977). The nomenclature in the first instance is striking because the genetics underlying this paper is called the neutral theory). Observe that in the limit $k \to \infty$, $\alpha \to 0$, $k\alpha \to \theta$, the Beta$(1+\alpha, (k-i)\alpha)$ random variables become Beta$(1, \theta)$ and (3.5) formally merges with (2.1).

The partition distribution of a sample from a population depends only on the unordered population frequencies and thus the partition distribution of a sample from $D(\alpha; k)$ is the same as that obtained from the urn $\mathcal{U}(\alpha; k)$. This indicates that the partition (1.1) of the limiting urn $\mathcal{U}$ should be the same as that obtained from the limiting population (2.1) (modulo some continuity arguments). This conclusion is correct and while the heuristic development leading to it is original the result is not new since it is known from the cited work of Watterson (1976) together with Kingman (1977) that the Ewens formula (1.1) describes a random sample taken from a Poisson–Dirichlet population with parameter θ (denoted by $PD(\theta)$ and described in the following paragraph). The size-biased permutation of $PD(\theta)$ is given by (2.1) (see McCloskey (1965), Engen (1975), Patil and Taillie (1977)) while the non-increasing order statistics of (2.1) are $PD(\theta)$. Both of them describe the same stochastic abundance structure and therefore give rise to the same partition distributions (1.1).

There are a number of characterizations of a Poisson–Dirichlet population (see Kingman (1978, Appendix)). The original definition and existence proof was given by Kingman (1975). He begins with a symmetric Dirichlet $D(\alpha; k)$ in which the population frequencies are re-arranged in decreasing order

$$P_{(1)} \geqslant P_{(2)} \geqslant \cdots \geqslant P_{(k)}$$

and he shows that for each fixed j, $(P_{(1)}, P_{(2)}, \ldots, P_{(j)})$ converge jointly to a vector $(P_1^*, \ldots, P_j^*)$ as $k \to \infty$, $\alpha \to \theta$, $k\alpha \to \theta$, and this vector is the jth joint marginal of a random probability $\underline{P}^*$ for which he coined the term Poisson–Dirichlet.

We note, finally, that $D(\alpha; k)$ does not converge to a proper random probability as $k\alpha \to \theta$. Each finite joint distribution tends to the zero vector since large k and small α realizes a population in which all alleles are present only in small proportions. Kingman's use of a permutation (descending order statistics) prior to passage to the limit overcomes this degeneracy. Similarly by first taking another (the size-biased) permutation we are again able to derive an appropriate limit. In fact, what the urn $\mathcal{U}$ is doing is directly generating the size-biased permutation of the Poisson–Dirichlet. We take this up in careful detail in the next section.

4. Relation to the Poisson–Dirichlet distribution

Watterson (1976) was the first to associate the Ewens sampling formula with a Poisson–Dirichlet population. His approach required the integration of a mixture of multinomials with a symmetric Dirichlet. Since the representation of the Poisson–Dirichlet as defined by Kingman loses the nice structure of the Dirichlet, which allowed the integration to be carried out, a similar calculation is not analytically feasible if the mixing distribution is Poisson–Dirichlet, and in fact

Watterson conjectures, "It is presumably the case, but not easy to prove, that [the Ewens formula] could be arrived at by directly sampling from the population described by $D(\alpha; k)$ rather than proceeding indirectly as we have done, letting k[the number of alleles] $\to \infty$".

A proof of this conjecture was provided by Kingman (1977) using weak convergence theory to graft limits of partitions and limits of populations. We sketch his arguments. Suppose given a sequence of populations $\{\underline{P}^{(k)}\}$ (Kingman assumes that the populations are finite and that samples are taken without replacement, but his argument is valid, and simplifies slightly, if sampling is with replacement, in which case the populations could be infinite, as is required here). Such a sequence is said to have the Poisson-Dirichlet limit if the population frequencies arranged in decreasing order converge in the sense of finite dimensional distributions to a Poisson-Dirichlet for some θ. This is shown to imply weak convergence of the corresponding induced distributions regarded as measures on an appropriate compact metric space. The partition distribution of a random sample is expressible as the expectation of a bounded continuous function on this space and hence if $\{\underline{P}^{(k)}\}$ have a $PD(\theta)$ limit then the partitions converge to the partition from a $PD(\theta)$ population. (We note in passing that Kingman also proves the converse assertion.) Watterson (1976) contains an explicit calculation that the partitions from $D(\alpha; k)$ converge to (1.1) as $\alpha \to 0$, $k \to \infty$ with $k\alpha \to \theta$, and (Kingman (1975)) by the very definition of $PD(\theta)$, the sequence $\{D(\alpha; k)\}$ has the $PD(\theta)$ limit. This identifies the Ewens formula as describing the partition from a Poisson-Dirichlet population.

This technical and indirect approach contrasts with the concreteness of Watterson's integrations for the symmetric Dirichlet.

A direct proof would be appropriate to exhibit for a result so basic but we have not seen one and we therefore provide such here. Our approach does not require determining the partition distribution, rather we compute the conditional probabilities of observing any particular type in the future given the current sample. This exposes the prominent role played by the Poisson-Dirichlet and its characteristic property which Ewens exploited, and which is a consequence of the representation (2.1).

Theorem 3 (Watterson, Kingman). *The partition distribution of a random sample $\{X_1, \dots, X_n\}$ from a Poisson-Dirichlet population $\underline{P}$ is given by Ewens' sampling formula.*

Proof. In view of Theorem A it suffices to establish that $\{X_1, X_2, \dots\}$ behaves sequentially (with appropriate relabelling) like the urn $\mathscr{U}$. We therefore need to evaluate the posterior probabilities, given the sample $\{X_1, X_2, \dots, X_n\}$, that the next observation will be novel or one of the types already observed (I thank Bruce Hill for a Bayesian interpretation which led to the step of conditioning on $\underline{P}$ below). Consider then $\Pr[X_{n+1} = X_j \mid X_1, X_2, \dots, X_n]$ where $1 \leq j \leq n$. By exchangeability of the sample it suffices to determine $\Pr[X_{n+1} = X_1 \mid X_1, \dots, X_n]$. We thus evaluate

$$\begin{aligned}\Pr[X_{n+1} = X_1 \mid X_1, \dots, X_n] &= E[\Pr[X_{n+1} = X_1 \mid X_1, \dots, X_n, \underline{P}] \mid X_1, \dots, X_n] \\ &= E[P_{X_1} \mid X_1, \dots, X_n].\end{aligned}$$

The random variable P_{X_1} is the first component in the size-biased permutation of $\underline{P}$ and is therefore Beta$(1, \theta)$. I claim that, more generally, the posterior distribution of P_{X_1} given $\{X_1, X_2, \ldots, X_n\}$ is Beta$(T, \theta+n-T)$ where $T = \#\{1 \leqslant j \leqslant n \colon X_j = X_1\}$. The argument for this, though quite straightforward, requires some extensive preparation. The key is the remarkable property of (2.1) that it defines a population which is invariant (in distribution) under size-biased permutation (Engen (1975)), from which ensues this implication (Hoppe (1986), Theorem 1(a)):

If a category if randomly (in proportion to its frequency) deleted from a $PD(\theta)$ population then the rescaled (to sum to unity) residual population is again $PD(\theta)$ and is independent of the frequency of the deleted category. (4.1)

The first observation X_1 by definition selects a category in proportion to its frequency. Give this category the new label #1. Its frequency P_{X_1} in the population will be denoted by both $P_1^{\#}$ and by $Z_1^{\#}$. The remaining categories are then relabelled #2, #,... with corresponding frequencies $P_2^{\#}, P_3^{\#}, \ldots$ in such a way that

$$P_2^{\#} = Z_2^{\#} \quad \text{and} \quad P_i^{\#} = Z_i^{\#} \prod_{j=2}^{i-1} (1 - Z_j^{\#}) \quad \text{for } i \geqslant 2 \tag{4.2}$$

where $\{Z_i^{\#}\}_{i=1}^{\infty}$ are independent and identically distributed Beta$(1, \theta)$ random variables. That this is possible is a direct consequence of (4.1) and the fact that (2.1) gives a representation of a $PD(\theta)$ population. The random sample is thus relabelled as $\{X_1^{\#}, X_2^{\#}, \ldots, X_n^{\#}\}$. Evidently $X_1^{\#} \equiv \#1$. For $j \geqslant 2$ either $X_j = X_1$, in which case $X_j^{\#} = \#1$, with

$$\Pr[X_j = X_1 \mid \underline{P}] = P_1^{\#}$$

or $X_j \neq X_1$ in which case $X_j^{\#} = \#i$ with (by (4.1) and (4.2))

$$\Pr[X_j^{\#} = \#i \mid X_j \neq X_1, \underline{P}] = P_i^{\#}$$

and thus

$$\Pr[X_j^{\#} = \#i \mid \underline{P}] = P_i^{\#}(1 - P_1^{\#}) = Z_i^{\#} \prod_{j=1}^{i-1} (1 - Z_j^{\#}) \quad \text{for } i \geqslant 2.$$

(Notice that the lower index in the product is now 1 in contrast with the 2 of (4.2).) Moreover the random variables $\{X_2^{\#}, X_3^{\#}, \ldots\}$ are still conditionally independent (given $\underline{P}^{\#}$) since relabelling does not alter their exchangeability, and represent a random sample from the same $PD(\theta)$ population. The posterior distribution P_{X_1} given $\{X_1, X_2, \ldots, X_n\}$ is in the relabelled # context, the same as the posterior distribution of $P_1^{\#}$ given a random sample $\{X_2^{\#}, \ldots, X_n^{\#}\}$ of size $n-1$ from $\underline{P}^{\#}$.

Thus assume that $\{X_1, X_2, \ldots, X_{n-1}\}$ is a random sample from a $PD(\theta)$ population $\underline{P}$ (the symbol # has been deleted for notational ease). For the evaluation of the posterior distribution P_1 given $\{X_1, X_2, \ldots, X_{n-1}\}$ we refer to Connor and Mosimann (1969) who have defined a generalized Dirichlet distribution $\underline{Q} = (Q_1, Q_2, \ldots, Q_m)$ by $Q_1 = U_1$, $Q_i = U_i \prod_{j=1}^{i-1} (1 - U_j)$, for $2 \leqslant i \leqslant m-1$, and $Q_m = 1 - \sum_{j=1}^{m-1} Q_j$ where $\{U_i\}$ are independent Beta(a_i, b_i) random variables. The $m-1$ dimensional vector $(Q_1, \ldots, Q_{m-1})$ has a joint density with respect to

Lebesgue measure given by

$$q_m^{b_{m-1}-1}\prod_{i=1}^{m-1}\frac{\Gamma(a_i+b_i)}{\Gamma(a_i)\Gamma(b_i)}q_i^{a_i-1}\left(\sum_{j=i}^{m}q_j\right)^{b_{i-1}-(a_i+b_i)}$$

on the simplex $q_i\geq 0$ and $q_1+\cdots+q_m=1$. Just as the posterior of a Dirichlet remains a Dirichlet with a change in the parameters the same is true of the generalized Dirichlet. In particular if a random sample is taken from $\underline{Q}$ and if n_i denotes the number of observations of type i (represented by frequency Q_i) then the posterior density of $(Q_1, Q_2,\ldots, Q_{m-1})$ is proportional to

$$q_m^{b_{m-1}-1}\prod_{i=1}^{m-1}q_i^{a_i-1}\left(\sum_{j=i}^{m}q_j\right)^{b_{i-1}-(a_i+b_i)}\prod_{i=1}^{m}q_i^{n_i}$$

which is a generalized Dirichlet with posterior U_i being Beta$(\bar{a}_i, \bar{b}_i)$ where $\bar{a}_i=a_i+n_i$, $\bar{b}_{m-1}=b_{m-1}+n_m$ and then recursively $\bar{b}_{i-1}=\bar{b}_i+n_i+b_{i-1}-b_i$. Note that b_0 is arbitrary, entering as a power 1^{b_0}.

Consider now our random sample $\{X_1, X_2,\ldots, X_{n-1}\}$ from $\underline{P}$ (described in the form (2.1)). The joint distribution is for fixed $(x_1, x_2,\ldots, x_{n-1})$

$$\Pr[X_1=x_1, X_2=x_2,\ldots, X_{n-1}=x_{n-1}, \underline{P}\in A]=\int_A \mu(dp)\prod_{i=1}^{\infty}p_i^{n_i}$$

where A is a Borel set in $[0,1]^{\infty}$, μ is the measure induced through the random mapping $\underline{P}$ on the Borel sets of $[0,1]^{\infty}$, and n_i is the number of $1\leq j\leq n-1$ such that $X_j=i$. (The product is only infinite in notation.) The posterior distribution of $\underline{P}$ is

$$\Pr[\underline{P}\in A\,|\,X_1=x_1,\ldots, X_{n-1}=x_{n-1}]=\int_A \mu(dp)\prod_{i=1}^{\infty}p_i^{n_i}\Big/\int_{[0,1]^{\infty}}\mu(dp)\prod_{i=1}^{\infty}p_i^{n_i}. \tag{4.3}$$

Those categories not represented among $\{x_1, x_2,\ldots, x_{n-1}\}$ are integrated out in the numerator (if A is suitable) and (4.3) for the specified $\{x_1, x_2,\ldots, x_{n-1}\}$ is then identical to the corresponding expression for the posterior of a random sample from the population

$$\left(P_1, P_2,\ldots, P_{m-1}, 1-\sum_{j=1}^{m-1}P_j\right)$$

where m is large enough to include all the observed categories (that is $m-1\geq\max\{x_1, x_2,\ldots, x_{n-1}\}$. The first $m-1$ dimensional marginal of this population is just the first $m-1$ dimensional marginal of (2.1) and is thus a generalized Dirichlet with parameters given by $a_i\equiv 1$ for all i and $b_i\equiv\theta$ for all i. The posterior distribution is therefore a generalized Dirichlet with $\bar{a}_i=a_i+n_i=1+n_i$ and solving recursively $\bar{b}_{i-1}=\bar{b}_i+n_i+b_{i-1}-b_i=\bar{b}_i+n_i=b_{m-1}+n_m+n_{m-1}+\cdots+n_i$. In particular $\bar{a}_1=n_1+1$ and $\bar{b}_1=\theta+n_m+\cdots+n_2=\theta+(n-1-n_1)$. But recall that $T=n_1+1$. Thus we identify the posterior of P_1 as being Beta$(\bar{a}_1, \theta+n-1-n_1)=$ Beta$(T, n+\theta-T)$ as claimed above and hence, with a return to the original notation in the statement of the theorem we are proving,

$$E[P_{X_1}|X_1,\ldots, X_n]=E[\text{Beta}(T, \theta+n-T)]=\frac{T}{\theta+n}. \tag{4.4}$$

If we define $t_i(n) = \#\{1 \leqslant j \leqslant n: X_j = X_i\}$ then by the exchangeability

$$\Pr[X_{n+1} = X_i \mid X_1, \ldots, X_n] = \frac{t_i(n)}{\theta + n} \tag{4.5}$$

and after summing over all distinct values among $\{X_1, \ldots, X_n\}$ we also get

$$\Pr[X_{n+1} \notin \{X_1, \ldots, X_n\} \mid X_1, \ldots, X_n] = \frac{\theta}{\theta + n}. \tag{4.6}$$

These are the conditional probabilities defining the urn process verifying that $\{X_1, \ldots, X_n\}$ has the Ewens partition structure and completing the proof.

Equation (4.6) brings us back full circle to Ewens' (1972) original paper describing how new alleles enter the sample. The expectation of (4.6) gives the unconditional probabilities

$$\Pr[n+1\text{th allele is the same as one of the first } n \text{ alleles}] = \frac{n}{\theta + n}$$

and

$$\Pr[n+1\text{th allele is novel}] = \frac{\theta}{\theta + n}.$$

In particular, as Ewens has pointed out, the number of different allelic types drawn on the first n draws has no bearing on the probability that on the $(n+1)$th draw a new type is obtained. Our proof thus has the nice feature of directly relating this sampling property of the allelic types to the structure of the underlying population (2.1).

Equation (4.4) implies that

$$\Pr[X_{n+1} = X_1 \mid X_1, \ldots, X_n] = P[X_{n+1} = X_1 \mid T]. \tag{4.7}$$

This is a statement of Johnson's sufficiency postulate (see Good (1965)). The symmetric Dirichlet $D(\alpha; k)$ also satisfies (4.7) and the right-hand side is then $(\alpha + T)/(k\alpha + n)$. If the population $\underline{P}$ from which the sample $\{X_1, X_2, \ldots, X_n\}$ is taken is, however, a mixture of Poisson-Dirichlets or symmetric Dirichlets then (4.7) does not hold since the data $\{X_1, \ldots, X_n\}$ provides information about θ or α which in turn is reflected in the posterior distribution of $\underline{P}$. In such mixture models the future probabilities of observing new alleles will depend on the multiplicities of all alleles in the current sample. The sampling theory for mixture models is therefore expected to be much more complicated and we suggest that attention should be restricted to those populations satisfying (4.7) as appropriate candidates for modelling purposes. It is thus of interest to find a general class of distributions for which (4.7) is particularly pleasing or tractable. In this connection we mention Hill (1979) who has proposed in a quite different context a class of models for species sampling. It is also worth pointing out that a non-symmetric Dirichlet $D(\alpha_1, \alpha_2, \ldots, \alpha_k)$ does not satisfy (4.7), rather it gives

$$\Pr[X_{n+1} = X_1 \mid X_1, \ldots, X_n] = (\alpha_{X_1} + T)/(\theta + n)$$

(where $\theta = \alpha_1 + \cdots + \alpha_k$) which is a function of both X_1 and T not just T. This consideration may be useful in the development of a sampling theory incorporating selection.

Both the Poisson-Dirichlet and the symmetric Dirichlet represent infinite populations satisfying Johnson's sufficiency postulate. To derive a class of finite populations satisfying (4.7) we can proceed as follows (although we use the Poisson-Dirichlet the same construction works for the symmetric Dirichlet). Let $\{X_1, X_2, \ldots, X_M\}$ represent a random sample from a Poisson-Dirichlet population $\underline{P}$. This sample determines a population $\underline{P}(M)$ of size M (use any convenient labelling, the actual choice being irrelevant for the discussion at hand).

Let $\{\xi_1, \xi_2, \ldots, \xi_n\}$, $n \leq M$ be a random sample without replacement taken from $\underline{P}(M)$. Since $\underline{P}(M)$ is itself determined by a random sample from $\underline{P}$ then by exchangeability $\{\xi_1, \xi_2, \ldots, \xi_n\}$ may be considered as a random sample from $\underline{P}$ (labels not withstanding). Consequently (4.5) and (4.6) are in force which in turn imply Ewens' formula, proving that a random sample of size n chosen without replacement from a population of size M whose partition is described by (1.1) also has the partition (1.1) (Theorem 7.1 of Kelly (1979) and Trajstman (1974)). The Poisson-Dirichlet can thus be interpreted as the infinite population analogue of finite populations which are described by Ewens' formula.

Actually if $\underline{P}$ is any population with a family of partition distributions induced by sampling and if a finite population $\underline{P}(M)$ is constructed from $\underline{P}$, as in the previous paragraph, then samples from $\underline{P}(M)$ also have the same partition distribution. Kingman (1978) uses this "consistency" condition to define partition structures but he deals with properties of samples while we are concerned with properties of populations. As is evident the two are hardly distinguishable.

5. Ages of alleles

In this section we establish a new methodology, based on the urn $\mathcal{U}$ for questions involving the ages of alleles in the infinite alleles models. We defer to Sect. 8 a rigorous validation based on mutation in the coalescent.

Recall that we have shown the following. If $\{X_1, X_2, \ldots\}$ represents a sequence of observations from a Poisson-Dirichlet population with parameter θ then the limiting proportions of the alleles indexed by their order of occurrence is given by (2.1). But Griffiths (unpublished) has shown that (2.1) also describes the proportions of the oldest, second oldest, ... alleles in the infinite alleles diffusion limit. Thus the distribution of types according to *ages in the population* is the same as the distribution of types according *to order in the sample.* One explanation for this has to do with size-biased sampling and reversibility. Using arguments based on the latter Kelly (1977) for the Moran model, and Watterson and Guess (1977) for the diffusion limit of the Wright-Fisher process show that the probability an allele is oldest in the population is its frequency. But this is also the probability of observing any particular, and hence the first allele (size-biased sampling). Thus the frequency in the population of the first observed allele is the same as the frequency of the oldest allele. Analogous arguments can be made for the second oldest, third oldest, etc. The use of size-biased sampling (permutation of labels) was also invoked by Hoppe (1986) to give an alternate proof of Kingman's (1978) sampling characterization of the Ewens formula. Another explanation based on genealogy will be given in Sect. 8.

This parallel between ages in the population and order in the sample will now be exploited to derive very readily results previouly determined by more laborious methods, as well as to shed some insight into why these are obtained. We have selected twelve examples for illustration below. Where there is both an (a) and (b) part, the former gives the known result while the latter presents the order approach and proof (where necessary). These show that the urn model provides a unified framework and powerful approach to diverse issues.

Let $\{X_1, X_2, \ldots\}$ be a sequence of observations from a Poisson–Dirichlet population $\underline{P}$. In order to obtain joint probabilities involving both the sample and the population, we make the identifications:

$\{X_1, \ldots, X_n\}$ represents the observed sample;

$\{X_{n+1}, X_{n+2}, \ldots\}$ represents the population.

(What we mean by the latter is that we can recover the population by the strong law of large numbers applied to $\{X_{n+1}, X_{n+2}, \ldots\}$.) In the sequel ε represents the type of a randomly selected allele from the sample, $S \equiv \#\{1 \leqslant j \leqslant n \colon X_j = \varepsilon\}$ is the number of alleles in the sample which are type ε, and $S_i = \#\{1 \leqslant j \leqslant n \colon X_j = X_i\}$ is the number of alleles of the same type as the ith observation in the sample. Observe that exchangeability forces S and S_i to be identically distributed.

1. (a) Watterson and Guess (1977). If an allele has frequency Y in the population then it is the oldest with probability Y (conditional on the population).

(b) If an allele has frequency Y in the population then it will be observed first in the sample with probability Y (conditional on the population).

2. (a) Watterson and Guess (1977); Kelly (1977). If an allele is represented by i individuals in a sample of size n then it is the oldest in the population with probability $i/(\theta+n)$.

(b) $\Pr[X_{n+1} = \varepsilon \mid S = i] = i/(\theta+n)$.

Proof. By (4.5) $\Pr[X_{n+1} = X_j \mid S_j = i] = i/(\theta+n)$. Hence

$$\Pr[X_{n+1} = \varepsilon \mid S = i] = \sum_{j=1}^{n} \frac{1}{n} \Pr[X_{n+1} = X_j \mid S_j = i] \frac{\Pr[S_j = i]}{\Pr[S = i]}$$

$$= i/(\theta+n).$$

3. (a) Watterson and Guess (1977); Kelly (1977). If an allele is represented by i individuals in a sample of size n, then it is the oldest in the sample with probability i/n.

(b) $\Pr[X_1 = \varepsilon \mid S = i] = i/n$.

This is immediate from exchangeability.

4. (a) Kelly (1977). In a sample of size n the oldest allele has i representatives with probability

$$\frac{\theta}{n}\binom{n}{i} \Big/ \binom{\theta+n-1}{i}.$$

(b) $\Pr[S_1=i]=\frac{\theta}{n}\binom{n}{i}\Big/\binom{\theta+n-1}{i}$.

Proof. From the urn model $\mathscr{U}$ the first allele observed carries the label 1. Its future occurrence in the sample can thus be described by a two-type Pólya urn indicating that $\Pr[S_1=i]$ is a mixture.

$$\Pr[S_1=i]=\int_{x=0}^{1}\binom{n-1}{i-1}x^{i-1}(1-x)^{n-i}\theta(1-x)^{\theta-1}\,dx \tag{5.1}$$

of binomials with success probability having Beta(1, θ) distribution (by (2.4) and (2.5)). This integrates to

$$\theta\binom{n-1}{i-1}\frac{\Gamma(i)\Gamma(n+\theta-i)}{\Gamma(n+\theta)}$$

which is another way of expressing the desired probability.

5. (a) Saunders, Tavaré, and Watterson (1984). If N_n is the number of types in the population which are older than the oldest allele in a sample of size n then

$$\Pr[N_n=k]=\frac{n}{\theta+n}\left(\frac{\theta}{\theta+n}\right)^k,\qquad k=0,1,\ldots.$$

(b) Let ν_n denote the number of types observed in the sequence $\{X_{n+1},X_{n+2},\ldots\}$ before a type in $\{X_1,X_2,\ldots,X_n\}$ is observed for the first time again. Then

$$\Pr[\nu_n=k]=\frac{n}{\theta+n}\left(\frac{\theta}{\theta+n}\right)^k,\qquad k=0,1,\ldots.$$

Proof. Let $\{T_1,T_2,\ldots\}$ denote those times $i\geqslant n+1$ for which X_i is either novel (not in the preceding set $\{X_1,X_2,\ldots,X_{i-1}\}$) or in $\{X_1,X_2,\ldots,X_n\}$.

$$\begin{aligned}\Pr[\nu_n\geqslant k]&=\Pr\left[\bigcap_{i=1}^{k}\{X_{T_i}\text{ is novel}\}\right]\\&=\prod_{i=1}^{k}\Pr\left[X_{T_i}\text{ is novel}\,|\,X_{T_j}\text{ is novel}(1\leqslant j\leqslant i-1)\right]\\&=\prod_{i=1}^{k}E[\Pr[X_{T_i}\text{ novel}\,|\,X_{T_j}\text{ novel}(1\leqslant j\leqslant i-1),\\&\qquad\qquad T_i]\,|\,X_{T_j}\text{ novel}(1\leqslant j\leqslant i-1)].\end{aligned}$$

Think of the $\{X_1,X_2,\ldots\}$ as generated by the urn $\mathscr{U}$. The conditioning specifies that the urn contains n "old" balls (representing the $\{X_1,X_2,\ldots,X_n\}$), one black ball (novel) and T_i-n-1 "recent" balls (the colours added subsequent to the nth drawing) just before the T_ith drawing and the T_ith drawing results in either an old ball or a black ball. Since old balls have unit mass while the black ball has mass θ it is apparent that the conditional probability of drawing a black ball, given that either a black ball or an old ball has been drawn, is $\theta/\theta+n$. Notice that T_i disappears and that this argument generalizes (4.6).

6. *Watterson (1974).* If a_i denotes the number of alleles represented i times in a sample of size n then

$$E[a_i]=\frac{\theta}{i}\binom{n}{i}\bigg/\binom{\theta+n-1}{i}.$$

Proof. Integrate $\Pr[S_1=i\,|\,a_1,\ldots,a_n]=ia_i/n$ obtaining $E[a_i]=n/i\Pr[S_1=i]$ and then use (5.1).

Observe that this method yields the interesting relationship

$$\Pr[\text{oldest allele in the sample is represented } i \text{ times}]=\frac{i}{n}E[a_i].$$

7. *Ewens (1973, 1979).* If K_n denotes the number of distinct alleles in a sample of size n then

$$E[a_i\,|\,K_n=k]=\frac{n!}{i(n-i)!}|S_{n-i}^{(k-1)}|/|S_n^{(k)}|$$

where $|S_n^{(k)}|$ is the coefficient of θ^k in $\theta(\theta+1)\cdots(\theta+n-1)$ (a Stirling number of the first kind).

Proof. $\Pr[S_1=i\,|\,a_1,\ldots,a_n,K_n]=ia_i/n$ and taking the conditional expectation with respect to K_n results in

$$E[a_i\,|\,K_n]=\frac{n}{i}\Pr[S_1=i\,|\,K_n].$$

According to Kingman's (1978) characterization theorem, after removal of the i alleles corresponding to the type of an individual chosen at random from the sample, the remaining $n-i$ alleles constitute a sample from the same population. Thus

$$\begin{aligned}\Pr[S_1=i\,|\,K_n=k]&=\Pr[S_1=i,K_n=k]/\Pr[K_n=k]\\&=\Pr[S_1=i]\Pr[K_{n-i}=k-1]/\Pr[K_n=k]\end{aligned}$$

and the assertion follows from (5.1) and Ewens' (1972) result $\Pr[K_n=k]=\theta^k|S_n^{(k)}|/\theta(\theta+1)\cdots(\theta+n-1)$.

Again observe that this method yields the relationship

$$\Pr[\text{oldest allele in the sample is represented } i \text{ times}\,|\,K_n]=\frac{i}{n}E[a_i\,|\,K_n]$$

and hence the following.

8. *Donnelly and Tavaré (1985).* In a sample of size n the probability that the oldest allele is represented by i individuals, given that there are k distinct alleles in the sample is

$$\frac{(n-1)!}{(n-i)!}|S_{n-i}^{(k-1)}|/|S_n^{(k)}|.$$

9. Ewens (1972).

$$\Pr[\text{two alleles drawn at random are of the same type}] \equiv E[\textstyle\sum P_i^2] = \frac{1}{1+\theta}.$$

Proof. Evaluate as $\Pr[X_1 = X_2]$ leading immediately to $1/(1+\theta)$.

10. Watterson and Guess (1977). Let $P_{(1)} = \max P_i$. Then for $0.5 \leqslant x \leqslant 1$ the density of $P_{(1)}$ is $\theta x^{-1}(1-x)^{\theta-1}$.

Proof. We use the notation in the proof of Theorem 1.

$$\begin{aligned}\Pr[P_{(1)} > x] &= \sum_{j=1}^{\infty} \Pr[P_j > x \text{ and } P_{(1)} = P_j]\\ &= \sum_{j=1}^{\infty} \Pr[\gamma_j > x \text{ and } P_{(1)} = \gamma_j]\\ &= \sum_{j=1}^{\infty} \Pr[\gamma_j > x]\\ &= \sum_{j=1}^{\infty} \Pr[\gamma_j > x \mid \lambda_j = 1] \Pr[\lambda_j = 1].\end{aligned}$$

Given $\lambda_j = 1$ then the limit proportion γ_j is Beta$(1, \theta+j-1)$ so the sum becomes

$$\begin{aligned}\sum_{j=1}^{\infty} \int_x^1 \frac{\Gamma(\theta+j)}{\Gamma(\theta+j-1)} (1-t)^{\theta+j-2}\, dt \frac{\theta}{\theta+j-1} &= \theta \int_x^1 \sum_{j=1}^{\infty} (1-t)^{\theta+j-2}\, dt\\ &= \theta \int_x^1 t^{-1}(1-t)^{\theta-1}\, dt.\end{aligned}$$

11. Ewens (1972). If $K_n = \#$ distinct alleles in a sample of size n then

$$\Pr[K_n = k] = \theta^k |S_n^{(k)}| / [\theta]^n.$$

Proof. The p.g.f. of K_n is (using the independence of the $\{\lambda_j\}$ in Theorem 1)

$$E[s^{K_n}] = \prod_{i=1}^{n} \left(\frac{i-1+\theta s}{i+\theta-1}\right)$$

the numerator of which generates the Stirling numbers above.

As a corollary, if t_i is the time of appearance of the ith novel allele then

$$\begin{aligned}\Pr[t_i = n] &= \Pr[K_{n-1} = i-1 \text{ and } \lambda_n = 1]\\ &= \frac{\theta^{i-1}|S_{n-1}^{(i-1)}|}{[\theta]^{n-1}} \frac{\theta}{n+\theta-1}\\ &= \frac{\theta^i |S_{n-1}^{(i-1)}|}{[\theta]^n}.\end{aligned}$$

12. Ewens (1972). The frequency spectrum $\phi(x)$ for the Poisson–Dirichlet population $\underline{P}$ is

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

Proof. Let f be a bounded measurable function on $[0, 1]$ and let $Q_1 = P_1^s$ be the first component in the size-biased permutation of $\underline{P}$. Then clearly

$$\sum f(P_i)P_i = E[f(Q_1) \mid \underline{P}]$$

and since Q_1 has a Beta$(1, \theta)$ density

$$\begin{aligned} E[\sum f(P_i)P_i] &= E[f(Q_1)] \\ &= \int_{x=0}^{1} f(x)\theta(1-x)^{\theta-1}\,dx. \end{aligned}$$

Choose for f the function

$$f(x) = \begin{cases} 1/x & \text{if } x > t \\ 0 & \text{if } x \leqslant t \end{cases}$$

to derive

$$E[\text{number of types whose frequency exceeds } t] = \int_{x=t}^{1} \theta x^{-1}(1-x)^{\theta-1}\,dx$$

showing that $\theta x^{-1}(1-x)^{\theta-1}$ is the frequency spectrum.

By setting $g(x) = xf(x)$ we can write

$$E[\sum g(P_i)] = \int_{x=0}^{1} g(x)\phi(x)\,dx.$$

This was obtained (for continuous f) by Kingman (1980), Eq. (3.4.2) by resorting to the Riesz representation theorem for positive linear functionals. Our argument seems more direct and relates the frequency spectrum to the size-biased permutation.

The multivariate frequency spectrum (Watterson (1974, 1976)) can also be derived using (in the bivariate case, for instance)

$$\sum_{i,j} f(P_i, P_j)\frac{P_iP_j}{1-P_i} = E[f(Q_1, Q_2) \mid \underline{P}]$$

where $Q_2 = P_2^s$ the second component in the size-biased permutation of $\underline{P}$.

6. A representation for Ewens' partition

A residual allocation model $\underline{Q} = (Q_1, Q_2, \ldots)$ (see Connor and Mosimann (1969), Patil and Taillie (1977)) is described in terms of an independent family $\{U_i\}$ of random variables called residual fractions. We may think of $\underline{Q}$ as defining how resources are assigned to a region or types (alleles, colours) are assigned to a population. The first colour is assigned to a proportion $U_1 \equiv Q_1$ of the population, the second colour to a proportion U_2 of the remaining (or residual) fraction $1 - U_1$ of the population and in general the ith colour is assigned to a proportion U_i of the residual fraction $\prod_{j=1}^{i-1}(1-U_j)$ left after the first $i-1$ colours have been assigned. In equation form

$$Q_1 = U_1, \qquad Q_i = U_i \prod_{j=1}^{i-1}(1-U_j), \qquad i \geqslant 2. \tag{6.1}$$

When there are but a finite number k of colours to be assigned (6.1) is defined only for $i \leqslant k-1$ and then $Q_k = \prod_{j=1}^{k-1}(1-U_j)$. This can be succinctly obtained by demanding that U_k be identically 1. From (2.1) we see that the size-biased permutation of the Poisson-Dirichlet $PD(\theta)$ is a residual allocation model with the residual fractions being identically distributed Beta$(1, \theta)$. Also the size-biased permutation (3.5) of the symmetric Dirichlet $D(\alpha; k)$ is a residual allocation model with the residual fractions U_i being Beta$(1+\alpha, (k-i)\alpha)$.

It was observed at the end of Sect. 4 that the Poisson-Dirichlet is the infinite population analogue of finite populations described by Ewens' partition. This raises the natural question of whether a corresponding representation holds for Ewens' partition.

Thus let $\underline{F}$ be a finite population of size M representing alleles whose unordered frequencies are distributed as in (1.1), but with M replacing n. For instance $\underline{F}$ might represent the Moran model in discrete time (Watterson (1974), Trajstman (1974)) at stationarity. Form the size-biased permutation $\underline{V}$ by selecting one individual at random and removing all V_1 alleles of the same type, then selecting another individual at random from the residual $M - V_1$ and removing all V_2 alleles of the same type, then continuing in this fashion until the entire population has been depleted (after K selections) and finally defining $\underline{V} = (V_1, V_2, \ldots, V_K)$.

Theorem 4. *$\underline{V}$ has the representation*

$$\begin{aligned} V_1 &= 1+\mathrm{Bin}(M-1, Z_1) \\ V_i &= 1+\mathrm{Bin}(M-1-V_1-\cdots-V_{i-1}, Z_i), \qquad 2 \leqslant i \leqslant K \end{aligned} \tag{6.2}$$

where $\{Z_i\}$ are i.i.d. Beta$(1, \theta)$ *random variables and K is the first integer such that $V_1 + V_2 + \cdots + V_K = M$.*

The notation Bin(r, Z) refers to a binomial random variable on r trials and success probability Z, both of which may be random.

Proof. According to Theorem 3, $\underline{F}$ may be realized as a random sample $\{X_1, \ldots, X_M\}$ from a Poisson-Dirichlet population $\underline{P}$. Because of exchangeability, selecting one observation at random from the set $\{X_1, \ldots, X_M\}$ is probabilistically equivalent to selecting the first observation X_1. The proof of Theorem 3 shows tha the urn $\mathscr{U}$ automatically generates the size-biased permutation of the partition determined by $\{X_1, \ldots, X_M\}$ and therefore $(V_1, \ldots, V_K)$ is equal in distribution to $(S_1(M), \ldots, S_K(M))$, using the notation of Sect. 2. But $S_1(M)$ can be described by a two-type Pólya urn with initial composition $(1, \theta)$ and then Eqs. (2.4) and (2.5) show that $S_1(M)$ is a Beta-binomial random variable of the specified form, verifying (6.2) for V_1. To get V_2 we then remove all $\{X_j: X_j = X_1\}$ before selecting another allele for the second component in the size-biased permutation. But those $X_j \neq X_1$ remaining after the first deletion represent a random sample from a population $\underline{Q}$ which is obtained by removal of one category at random from $\underline{P}$. Now $\underline{P}$ is invariant under this operation (size-biased permutation), meaning that $\underline{Q}$ is also Poisson-Dirichlet, and moreover $\underline{Q}$ is independent of the frequency of the deleted category (see Hoppe (1986)). Thus $S_2(M)$ is also a Beta-binomial with the mixing variable Z_2 independent of Z_1. This argument works for $S_i(M)$ in general establishing (6.2).

Corollary. *The population* $\underline{Y}$ *described by* (6.2) *is invariant under size-biased permutation.*

It is implicit in the description of a residual allocation model that the population be a continuum. To come up with an analogous concept for discrete populations it is necessary to rethink the independence structure of the residual fractions. Using (6.2) as our guide we therefore focus not on the proportions assigned to the ith type but rather the actual amount assigned, which, for the model (6.1) is $U_i \prod_{j=1}^{i-1}(1-U_j)$. This quantity depends on $\{U_1, U_2, \ldots, U_{i-1}\}$ only through $\prod_{j=1}^{i-1}(1-U_j)$, which is the remaining portion of the population. This observation suggests an appropriate reformulation of the residual allocation model for a discrete population.

Suppose a discrete population of size $M<\infty$ is to be painted with colours 1, $2, \ldots, C$ where C may be finite or infinite. We present a method for assigning the colours sequentially in such a way that once the first $i-1$ colours have been assigned then the ith is assigned by a randomized procedure which depends on the previous assignments only through their cumulative total. Specifically let $\{R_i(n)\colon 0 \leq n \leq M\}$ for each $1 \leq i \leq C$ be a family of integer-valued random variables where $0 \leq R_i(n) \leq n$. The families are independent but we make no assumptions about dependence within a family. If colours $1, 2, \ldots, i-1$ have been assigned to $n_1, n_2, \ldots, n_{i-1}$ individuals respectively then

$$n_i = R_i\left(M - \sum_{j=1}^{i-1} n_j\right)$$

of the remaining individuals are assigned colour i. This process is continued until the population is exhausted with colour K (random) and the vector $(n_1, n_2, \ldots, n_K)$ is said to be a discrete residual allocation model.

Comparing with (6.2) we see that the Ewens partition has a representation as a discrete residual allocation model with residual functions

$$R_i(n) = 1 + \mathrm{Bin}(n-1, Z_i)$$

where $\{Z_i\}$ are independent Beta$(1, \theta)$, thereby providing affirmation of the question raised in the second paragraph above.

The construction typified by Theorem 4 works for samples from any (continuous) residual allocation model $\underline{Q}$ as described by (6.1). Suppose a random sample (with replacement) of size M is drawn from $\underline{Q}$. Let n_i be the number of observations in category (colour) i while K is the first index such that $n_1 + n_2 + \cdots + n_K = M$.

Theorem 5. $(n_1, n_2, \ldots, n_K)$ *is a discrete residual allocation model with residual functions* $R_i(n) = \mathrm{Bin}(n, U_i)$.

Proof. The assertion of this theorem is that

$$n_i = \mathrm{Bin}\left(M - \sum_{j=1}^{i-1} n_j,\ U_i\right). \tag{6.3}$$

Now by definition, for arbitrary non-negative integers $x_1, x_2, \ldots, x_r$ summing to M

$$\Pr[n_1 = x_1, n_2 = x_2, \ldots, n_r = x_r] = E\left[\binom{M}{\underline{x}} \prod_{i=1}^{r} Q_i^{x_i}\right]$$

where $\binom{M}{x}$ is the multinomial coefficient $M!/\prod_{i=1}^{r} x_i!$ On the other hand, according to (6.3)

$$\Pr[n_1=x_1, n_2=x_2, \dots, n_r=x_r]=E\left[\prod_{i=1}^{r}\binom{M-y_{i-1}}{x_i}U_i^{x_i}(1-U_i)^{M-y_i}\right]$$

where $y_i=x_1+\cdots+x_i$ and $y_0=0$. The multinomial coefficients in the two expressions are equal and in the first one U_i appears with a power x_i while $(1-U_i)$ appears with a power $x_{i+1}+x_{i+2}+\cdots+x_M=M-x_1-\cdots-x_i=M-y_i$. Hence the two expressions are the same, proving the theorem.

This proof may fail to expose the nice way in which samples from a residual allocation model may be generated by successive sweeps through an array of M cells, picking out which colours go to which cells. Given U_1 carry out independent Bernoulli trials, with success probability U_1, at each of the M cells. A success assigns the colour 1 to the cell and it drops out. For the remaining cells again carry out independent Bernoulli trials, this time with success probability U_2. A success assigns colour 2 and this "sweeeping through" the cells with successive Bernoulli trials is repeated until all cells are assigned colours. The contents of cell i are identified with the ith observation from $\underline{Q}$.

When $\underline{Q}$ is the Poisson-Dirichlet then (6.3) gives a representation different from (6.2). There is no contradiction. The random variables $\{V_i\}$ in (6.2) are already in size-biased form, while the $\{n_i\}$ in (6.3) are not. Observe also that some of the $\{n_i\}$ may be zero while the $\{V_i\colon 1\leqslant i\leqslant K\}$ are all positive.

Several methods have been proposed for simulating a random sample from a Poisson-Dirichlet family, notably we mention Stewart's appendix to Fuerst et al. (1977), Griffiths and Li (1983), and Watterson (1985). Equation (6.2) provides another way based on simulating Beta-binomial random variables. Additionally it generates the age classes themselves directly. It should be possible to generate samples very efficiently this way.

Finally we use (6.2) to derive the joint distribution of the $\{V_i\}$. First we need a typical term

$$\begin{aligned}\Pr[S_1(M)=i]&=\int_{x=0}^{1}\binom{M-1}{i-1}x^{i-1}(1-x)^{M-i}\theta(1-x)^{\theta-1}\,dx\\&=\theta\binom{M-1}{i-1}\frac{\Gamma(i)\Gamma(M+\theta-i)}{\Gamma(M+\theta)}.\end{aligned}\tag{6.4}$$

(We computed this earlier as (5.1).) Thus

$$\begin{aligned}\Pr[V_1=x_1,\dots,V_r=x_r,K\geqslant r]&=\Pr[S_1(M)=x_1,\dots,S_r(M)=x_r,K\geqslant r]\\&=\prod_{i=1}^{r}\binom{M-1-y_{i-1}}{x_i-1}\\&\quad\times\int_{t=0}^{1}t^{x_i-1}(1-t)^{M-y_{i-1}-x_i}\theta(1-t)^{\theta-1}\,dt\\&=\prod_{i=1}^{r}\theta\binom{M-y_{i-1}-1}{x_i-1}\frac{\Gamma(x_i)\Gamma(M+\theta-y_i).}{\Gamma(M+\theta-y_{i-1})}\end{aligned}$$

by (6.4) where $y_i = x_1 + \cdots + x_i$ and $y_0 = 0$. This expression simplifies to

$$\frac{\theta^r M!\,\Gamma(M+\theta-y_r)}{(M-y_r)!\,\Gamma(M+\theta)} \prod_{i=1}^{r} \frac{1}{M-y_{i-1}}. \tag{6.5}$$

The evaluation of

$$\Pr[V_1 = x_1, \ldots, V_r = x_r, K = r]$$

proceeds identically but with the side constraint $y_r = M$ giving

$$\frac{\theta^r M!\,\Gamma(\theta)}{\Gamma(M+\theta)} \prod_{i=1}^{r} \frac{1}{t_i} \tag{6.6}$$

where $t_i = M - y_{i-1} = x_i + x_{i+1} + \cdots + x_M$. Equations (6.5) and (6.6) have recently been derived by Donnelly and Tavaré (1985), by entirely different methods based on a coalescent, as the joint distribution of the age partition in a sample from the infinite alleles model, thus providing another manifestation of the identification of order with age. (We proved a special case $r = 1$ in Sect. 5.)

We close by observing that if the equations in (6.2) are divided by M and then $M \to \infty$ it follows immediately by the strong law of large numbers for binomial random variables that

$$\left(\frac{V_1}{M}, \frac{V_2}{M}, \ldots, \frac{V_r}{M}\right)$$

converges almost surely for each fixed r as $M \to \infty$ to $(Z_1, Z_2(1 - Z_1), \ldots, Z_r \prod_{i=1}^{r-1} (1 - Z_i))$. This is not really a proof of Theorems 1 and 2 because (6.2) depended on these theorems, but the derivation is very suggestive that (6.2) can be used to model a multitude of finite populations whose sampling properties in the limit of large population size are described by the Ewens sampling formula (Kingman (1977)).

7. The reversed chain as a partition structure

According to Kingman (1978) a partition structure is a family $\{P_n\}$ of distributions on partitions satisfying the following consistency relation: If a sample of size n has allelic partition distribution P_n then the distribution of a random subsample of size m taken from this sample is P_m.

The relationship between P_m and P_n is expressed by

$$P_m = \sigma_{mn} P_n \tag{7.1}$$

where $\{\sigma_{mn}\}$ is a family of linear transformations satisfying

$$\sigma_{ln} = \sigma_{lm}\sigma_{mn} \qquad (l < m < n) \tag{7.2}$$

which in the special case $n = m + 1$ reduces to

$$P_m(a_1, \ldots, a_m) = \frac{a_1+1}{m+1} P_{m+1}(a_1+1, a_2, \ldots, a_m, 0) + \sum_{i=2}^{m+1} \frac{i(a_i+1)}{m+1} P_{m+1}(a_1, \ldots, a_{i-1}-1, a_i+1, \ldots). \tag{7.3}$$

The Ewens distribution is a partition structure arising naturally through a process $\{\Pi_n\}$ thus directing us to examine (7.1) and (7.2) in the context of Markov chains.

The process $\{\Pi_n\}$ of Theorem A begins at $\Pi_0 = \emptyset$, the empty partition. This is a consequence of the initial composition of the urn $\mathcal{U}$. We can, however, impose an initial composition prescribed by a measure Π_0 to define a more general Markov chain with one-step transition probabilities

$$P(a, b) = \Pr[\Pi_{n+1} = b \,|\, \Pi_n = a]$$

from partition $a \equiv (a_1, \ldots, a_r)$ of the integer r to partition b (necessarily of the integer $r+1$), and arbitrary initial distribution Π_0, where

$$P(a, b) = \begin{cases} \dfrac{\theta}{\theta + r} & \text{if } b = (a_1 + 1, a_2, \ldots, a_r, 0) \\ \dfrac{i a_i}{\theta + r} & \text{if } b = (a_1, \ldots, a_i - 1, a_{i+1} + 1, \ldots, a_r, 0) \\ \dfrac{r a_r}{\theta + r} & \text{if } b = (a_1, \ldots, a_r - 1, 1). \end{cases} \tag{7.4}$$

We next consider the time-reversed probabilities

$$\Pr[\Pi_n = a \,|\, \Pi_{n+1} = b] = \Pr[\Pi_{n+1} = b \,|\, \Pi_n = a] \Pr[\Pi_n = a] / \Pr[\Pi_{n+1} = b].$$

These will depend on the initial state Π_0 but when $\Pi_0 = \emptyset$ it may be verified using (1.1) that

$$\Pr[\Pi_n = a \,|\, \Pi_{n+1} = b] = \begin{cases} \dfrac{a_1 + 1}{n+1} & \text{if } b = (a_1 + 1, a_2, \ldots, a_n, 0) \\ \dfrac{(r+1)(a_{r+1} + 1)}{n+1} & \text{if } b = (a_1, \ldots, a_r - 1, a_{r+1} + 1, \ldots, a_n, 0) \\ 1 & \text{if } b = (a_1, \ldots, a_n - 1, 1). \end{cases} \tag{7.5}$$

Denote the matrix of reversed transition probabilities by Σ_{nn+1}. Introduce the notation $T_n(a) = \Pr[\Pi_n = a \,|\, \Pi_0 = \emptyset]$ and compute $T_n(a) = \sum_b \Pr[\Pi_n = a \,|\, \Pi_{n+1} = b] T_{n+1}(b)$ which simplifies to

$$\begin{aligned} T_n(a_1, \ldots, a_n) = {} & \frac{a_1 + 1}{n+1} T_{n+1}(a_1 + 1, a_2, \ldots, a_n, 0) \\ & + \sum_{r=2}^{n} \frac{r(a_r + 1)}{n+1} T_{n+1}(a_1, \ldots, a_{r-1} - 1, a_r + 1, \ldots, a_n, 0) \\ & + T_{n+1}(a_1, \ldots, a_n - 1, 1). \end{aligned}$$

This is (7.3). We may in the same fashion for $l < m < n$ compute $\Pr[\Pi_l = a \,|\, \Pi_n = b]$ by a decomposition based on Π_m. If Σ_{mn} is the matrix with components $\Sigma_{mn}(a, b) = \Pr[\Pi_m = a \,|\, \Pi_n = b]$ then we find

$$T_m = \Sigma_{mn} T_n$$

and

$$\Sigma_{ln} = \Sigma_{lm} \Sigma_{mn}.$$

This shows that (7.1) and (7.2) are the Chapman-Kolmogorov equations for the time-reversed Markov chain of partitions. This reversed chain begins in some partition a and passes through partitions of decreasing integers. Further elucidation and interpretation of the reverse process will be made in the next section using the coalescent with mutation where it will be shown that the reverse process follows the lines of descent back in time through successive generations thereby describing the genealogy.

8. The urn and genealogy of the coalescent with mutation

Kingman (1982a, b, c) has introduced a continuous time Markov chain $\mathcal{R}_t$, the n-coalescent, taking values in $\mathcal{E}_n$, the set of all equivalence relations on $\{1, 2, \ldots, n\}$, to describe the common ancestry among n individuals randomly selected from a haploid population. Two individuals are in the same equivalence class at time t if and only if they share a common ancestor t time units in the past. The coalescent traces back the genealogy of a sample of individuals and provides a richer sample space structure than the classical methods (such as diffusion approximation) which keep track only of the numbers of each type in successive generations.

A sample path may be visualized as an inverted tree rooted in n vertices, each corresponding to one individual and numbered with the integers $\{1, 2, \ldots, n\}$, from which emanate vertical branches (lines of descent) passing backward through genealogical time (but forward in the coalescent). At random times, two branches meet at a vertex (representing a common ancestor) and then continue as one branch up the line of descent. The rules governing the times and choices of branch meetings guarantee that $\mathcal{R}_t$ is Markov. Specifically, in the infinitesimal time interval $(t, t+h)$ the probability is $h+o(h)$ that any two branches will meet. If there are i branches then the overall probability of a change in state is $\frac{1}{2}i(i-1)h+o(h)$ and the selection of branches to coalesce is made completely at random. The process $\mathcal{R}_t$ is then obtained from a horizontal cross-section through the tree at time t by identifying those individuals descending from each intersected branch as belonging to the same equivalence class. There is associated with $\mathcal{R}_t$ a pure death process D_t where $D_t = |\mathcal{R}_t|$, $|\cdot|$ denoting the number of equivalence classes. D_t is the number of intersected lines of descent at time t and therefore also equals the number of distinct ancestors, at time t in the past, of the sample (taken at time 0).

$\mathcal{R}_t$ passes through a sequence

$$R_n, R_{n-1}, \ldots, R_1 \tag{8.1}$$

of equivalence relations on $\{1, 2, \ldots, n\}$ spending an exponential amount of time with mean $2/k(k-1)$ in state R_k. Here $R_n = \{(i, i)\colon 1 \leq i \leq n\}$ and $R_1 = \{(i, j)\colon 1 \leq i, j \leq n\}$. The sequence (8.1) has the property that R_{i-1} is obtained by combining two equivalence classes in R_i and thus the number of equivalence classes is $|R_i| = i$.

Kingman (1982c) derives Ewens formula as a consequence of mutation in the coalescent. In particular suppose that each equivalence class in $\mathcal{R}_t$ is subject to a mutation in the interval $(t, t+h)$ with probability $(\theta/2)h+o(h)$. Then define a random equivalence relation $\mathcal{R}$ on $\{1, 2, \ldots, n\}$ by grouping individuals as follows. Two individuals are in the same random equivalence class if no mutation occurs up either's line of descent to their first common ancestor. Kingman (1982c, Theorem 4) shows that if $\xi \in \mathscr{E}_n$ has k equivalence classes with sizes $\lambda_1, \lambda_2, \dots, \lambda_k$ then

$$\Pr[\mathscr{R} = \xi] = \frac{\theta^k \prod_{j=1}^{k} (\lambda_j - 1)!}{[\theta]^n}. \tag{8.2}$$

If (8.2) is multiplied by the combinatorial term counting the number of equivalence relations with a given partition then (1.1) arises.

Now (8.2) also occurs in Hoppe (1984, Eq. (2)) with reference to the urn process $\{X_1, X_2, \dots, X_n\}$ described in Sect. 1 above. Specifically

$$\Pr[X_1 = x_1, \dots, X_n = x_n] = \frac{\theta^k \prod_{j=1}^{k} (S_j(n) - 1)!}{[\theta]^n}. \tag{8.3}$$

In fact there is a one-to-one correspondence which we now describe, between sequences $\{x_1, x_2, \dots, x_n\}$ of urn paths and equivalence relations $\xi \in \mathscr{E}_n$. In one direction, given $\{x_1, x_2, \dots, x_n\}$ define an equivalence relation ξ by placing i and j in the same class iff $x_i = x_j$. Conversely given ξ set $x_i \equiv x_i(\xi) = 1$ if and only if $(1, i) \in \xi$. Let $j(2) = \min\{i\colon x_i \neq 1\}$ and put $x_i = 2$ if and only if $(j(2), i) \in \xi$. Then let $j(3) = \min\{i\colon x_i \neq 1 \text{ or } 2\}$ and continue in the obvious way to construct a path $\{x_1, \dots, x_n\}$. The urn process $\{X_1, \dots, X_n\}$ is associated in this fashion with a random equivalence relation and by (8.2) and (8.3) this random equivalence relation has the same distribution as the $\mathscr{R}$ of Kingman.

There is a deep connection between the urn process $\mathscr{U}$ and the genealogy of the sample, to which we now turn, to explain why (8.2) and (8.3) are identical. Recall that $\mathscr{R}_t$ tracks the genealogy of a *fixed* sample of size n, and (8.2) describes the distribution at time 0 of the random equivalence relation $\mathscr{R}$ obtained from $\mathscr{R}_t$. This is a probability on the set $\mathscr{E}_n$ of all equivalence relations defined on $\{1, 2, \dots, n\}$. $\mathscr{U}$ however defines a process of partitions on sets whose sizes are *changing*. In particular Π_n defines a partition of the integer n, Π_{n-1} a partition of the integer $n-1, \dots, \Pi_1$ a partition of one. The equivalence relation $\mathscr{R}$ induces a partition $a = (a_1, a_2, \dots, a_n)$ of the integer n where a_i is the number of equivalence classes in $\mathscr{R}$ having size i. In view of Kingman's result that this partition has the Ewens sampling distribution, namely Π_n, this suggests that we construct a process of random equivalence relations on sets of decreasing sizes whose induced partition process has the same distribution as the reverse chain $\{\Pi_n, \Pi_{n-1}, \dots, \Pi_1\}$. For this construction we need candidates for the sets upon which these random equivalence relations will be defined. While $\mathscr{R}$ is defined on the original sample of size n it will ensue that the appropriate sets to use comprise decreasing lines of descents which have not undergone mutation from a given time in the past to the present.

If a time t is fixed then each member of the sample traces back to one of the D_t ancestors alive at time t in the past. Some are mutations from these D_t ancestors while others are not. Denote by L_t the number of ancestors at time t who have non-mutated descendents in the sample. These we call non-mutant ancestors and a line of descent from such an ancestor to the present is called an original line of descent. (We tacitly assume that mutations are therefore not included in the line of descent of their parents but are considered to begin new lines of descent (Griffiths (1980)).)

Observe that L_t is a pure death process, changes of state resulting from either coalescence or mutation. Coalescence occurs at unit rate to each pair of original lines of descent giving an overall rate $i(i-1)/2$ when $L_t = i$. Mutations occur at rate $\theta/2$ along each line and thus the overall mutation rate is $i\theta/2$. Combining we find that L_t changes from i to $i-1$ in time $(t, t+h)$ with probability $(i(\theta+i-1)/2)h+o(h)$. Denote by $T_n < T_{n-1} < \cdots < T_1$ the times at which L_t changes state and for definiteness set $L_t = i-1$ at T_i making L_t right-continuous.

Each member of the sample either traces back without mutation to one of the L_t non-mutant ancestors, or else is a mutation from one of the D_t ancestors. Those descending without mutation are the same type as their ancestor and it is natural then to trace back farther in time the history of these L_t non-mutant ancestors to define an equivalence relation (akin to $\mathcal{R}$) on the L_t original lines of descent which will group into equivalence classes non-mutant ancestors of the same type. If this is done for each t then the resulting process should trace the changing genealogy of the sample into the past.

These considerations motivate the definition of a process $\mathcal{S}_t$ of equivalence relations on the L_t original lines of descent which places any two lines into the same equivalence class of $\mathcal{S}_t$ if and only if upon following the ancestors they represent farther up their lines of descent (that is farther into the past) there is no mutation to either before their first common ancestor (that is in (t, T) where T is the moment they first coalesce). This specification mimics that of $\mathcal{R}$ but respective the L_t non-mutated ancestors at each time t rather than the individuals in the sample (who are the non-mutated ancestors at time $t=0$).

As does $\mathcal{R}_t$ so does $\mathcal{S}_t$ pass through a sequence

$$S_n, S_{n-1}, \ldots, S_1$$

of equivalence relations. In contrast, though, S_i is an equivalence relation on a set of cardinality i, not on the same fixed set $\{1, 2, \ldots, n\}$, and of course, it is generally not the case that $|S_i| = i$.

When L_t decreases by one $\mathcal{S}_t$ changes state. If by mutation then $\mathcal{S}_t$ loses one equivalence class of cardinality one and $|\mathcal{S}_t|$ decreases by one. Else, a coalescence results in an equivalence class, of cardinality greater than one, losing a member but $|\mathcal{S}_t|$ then of course stays the same. Whether by mutation or coalescence the transitions from S_i to S_{i-1} are caused by the loss of an original line of descent.

The process $\{S_i\}_{i=n}^{1}$ provides a sequence of random equivalence relations on sets of decreasing size $n, n-1, \ldots, 1$ paralleling the reverse partition chain determined by $\mathcal{U}$. The marginal distribution of each S_i is easy to find. Just as Kingman determined the distribution (8.2) of $\mathcal{R}$ (which is just our S_n) we can use a backward Kolmogorov argument to prove that

$$\Pr[S_i = \xi] = \frac{\theta^k \prod_{j=1}^{k} (\lambda_j - 1)!}{[\theta]^i} \tag{8.4}$$

where $\{\lambda_1, \lambda_2, \ldots, \lambda_k\}$ are the sizes of the equivalence classes of ξ. After all, in view of the Poisson nature of coalescence and mutation the i original lines of descent may be looked upon as representing a sample of size i defining an i-coalescent from which (8.4) is a recontexted version of (8.2).

While the changes of state in $\mathcal{R}_t$ consequent from the joining of two equivalence classes the behaviour of $\mathcal{S}_t$ is determined somewhat differently. Kingman (1982c, Theorem 1) shows that the death process D_t and the jump chain $\{R_k;\ k=n, n-1, \ldots, 1\}$ are independent and by definition $\mathcal{R}_t = R_{D_t}$. In the case at hand we have a death process L_t and a jump chain $\{S_k;\ k=n, n-1, \ldots, 1\}$ with $\mathcal{S}_t = S_{L_t}$. However L_t and $\{S_k\}$ are not independent since S_k "looks into the future" by tracing lines of descent until their coalescence and searching for mutations. For example if $L_t = i$

$$\Pr[L_{t+\Delta t} = i-1 \mid L_t = i] = \tfrac{1}{2}i(i+\theta-1)\Delta t + o(\Delta t)$$

but if we were to, additionally, condition $\mathcal{S}_t$ and for simplicity let $\mathcal{S}_t$ be the equivalence relation $\{(l, m);\ 1 \leq l, m \leq i\} \equiv \xi$ corresponding to the event that all of the i original lines of descent at time t trace back without mutation to a single ancestor, then given $L_t = i$ and $\mathcal{S}_t = \xi$ the next original line of descent must be lost by coalescence since the event $\{\mathcal{S}_t = \xi\}$ forbids mutations. In view of the "competing Poissons" (coalescence versus mutation)

$$\Pr[L_{t+\Delta t} = i-1 \mid L_t = i, \mathcal{S}_t = \xi] = \tfrac{1}{2}i(i-1)\Delta t + o(\Delta t)$$

displaying the lack of independence between L_t and $\{S_k\}$.

We have deliberately ignored numbering the lines of descent in the process L_t. Members of the original sample of size n can be numbered in some fashion (usually random) using the integers $\{1, 2, \ldots, n\}$, but when two lines of descent join there is no natural or best number for the line emerging from their common ancestor. More importantly the probabilistic structure is not dependent on the choice (which may also obscure the simplicity of the process). In particular while the marginal distributions of $\{S_i\}$ do not depend on the choice (the distribution of S_i depends only on the sizes of its equivalence classes) the joint and hence conditional distributions do.

To demonstrate this dependence of the joint distributions, consider the case where $n = 3$ and members of the initial sample of size 3 are numbered in some fashion with the integers 1, 2, 3. Suppose the backward genealogy has a coalescence at T_3 between 1 and 2, then a mutation at T_2 to the line 3 (numbered in the initial sample) and finally a mutation at T_1 to the line resulting from the joining of 1 and 2. Suppose the numbering rule is as follows. A mutation event occurring to an original line of descent eliminates its number. If a coalescence event occurs between two original lines of descent then the choice for the line emanating from their joining will be the smaller of the two numbers. In this example the equivalence classes in S_3 are $\{1, 2\}$ and $\{3\}$, the equivalence classes in S_2 are $\{1\}$ and $\{3\}$ and the equivalence class in S_1 is $\{1\}$.

Let η be the equivalence relation with classes $\{1\}$ and $\{3\}$. Suppose given $S_2 = \eta$. Observe that there are only two choices for S_3:

ξ_1 with equivalence classes $\{1\}$, $\{2\}$, and $\{3\}$ corresponding to a mutation at T_3;

ξ_2 with equivalence classes $\{1, 2\}$ and $\{3\}$ corresponding to a coalescence at T_3.

The relation ξ with equivalence classes $\{1\}$ and $\{2, 3\}$ is not permitted since the joint event $\{S_2 = \eta$ and $S_3 = \xi\}$ would require a coalescence between lines 2 and 3 in ξ in which case the line emerging from their coalescence would be numbered with 2 (being the lesser of 2 and 3) forcing S_2 to have classes $\{1\}$ and $\{2\}$, a contradiction. Hence

$$\Pr[S_3 = \xi \mid S_2 = \eta] = 0.$$

It is evident that a different numbering can be used to violate this equation for the same choices of ξ and η.

Accordingly it is preferable to let $\mathscr{S}_t$ (resp. S_i, $1 \leq i \leq n$) prescribe a partition process $\mathscr{A}_t$ (resp. A_i) which is the partition of the integer L_t (resp. i) determined by the sizes of the equivalence classes in $\mathscr{S}_t$ (resp. S_i). Thus $\mathscr{A}_t$ is a continuous time process passing through the states

$$A_n, A_{n-1}, \ldots, A_1.$$

Theorem 6. *$\{\mathscr{A}_t\}$ is a Markov process whose embedded reverse jump process $\{A_1, A_2, \ldots, A_n\}$ is equivalent to the partition chain $\{\Pi_1, \Pi_2, \ldots, \Pi_n\}$ of Sect* 1.

Proof. The process $\{\mathscr{A}_t\}$ looks into the future of the coalescent and to verify the Markov property it is convenient to reverse time turning the future into the past. Thus fix a reference time $s > 0$ and define the reverse process

$$\mathscr{A}_t^- = \mathscr{A}_{s-t}, \qquad 0 \leq t \leq s.$$

The past σ-algebra $\sigma\{\mathscr{A}_u^-: 0 \leq u \leq t_0\}$ at time t_0 is $\sigma\{\mathscr{A}_v: s - t_0 \leq v \leq s\}$, which depends only on the number of original lines of descent at time $s - t_0$ and the mutation-coalescence beyond $s - t_0$. Now let $t_1 > t_0$ and consider the conditional distribution of $\mathscr{A}_{t_1}^-$ given $\sigma\{\mathscr{A}_u^-: 0 \leq u \leq t_0\}$. This is the same as the conditional distribution of $\mathscr{A}_{s-t_1}$ given $\sigma\{\mathscr{A}_v: s - t_0 \leq v \leq s\}$. $\mathscr{A}_{s-t_1}$ is determined by the mutation-coalescence in $(s - t_1, s - t_0)$ together with the mutation-coalescence subsequent to $s - t_0$ the latter of which is completely captured by $\mathscr{A}_{s-t_0}$. Hence the conditional distribution of $\mathscr{A}_{s-t_1}$ given $\sigma\{\mathscr{A}_v: s - t_0 \leq v \leq s\}$ is the same as the conditional distribution of $\mathscr{A}_{s-t_1}$ given $\mathscr{A}_{s-t_0}$ which shows that $\{\mathscr{A}_t^-\}$ is Markov on $0 \leq t \leq s$. Since the Markov property is preserved under time reversal (a symmetric version of the Markov property asserts that the future and the past are conditionally independent given the present) it follows that $\{\mathscr{A}_t\}$ is Markov on $0 \leq t \leq s$, and s being arbitrary the first part of the theorem follows.

Because $\{\mathscr{A}_t\}$ is Markov so is the embedded jump process $\{A_n, A_{n-1}, \ldots, A_1\}$ and then also its time reversal $\{A_1, A_2, \ldots, A_n\}$ whose one-step transition probabilities we now calculate. First we consider $\Pr[A_n = b \mid A_{n-1} = a]$ for partitions a and b of $n - 1$ and n respectively. There are two cases.

(1) $a_{r-1} = b_{r-1} + 1$, $a_r = b_r - 1$, the remaining $a_j = b_j$ $(j \leq n - 1)$: This occurs when one of the equivalence classes in S_n necessarily having cardinality $r \geq 2$ is reduced by one because of a coalescence at T_n involving two of the n original lines of descent. The joint probability

$$\Pr[A_n = b, A_{n-1} = a]$$

may be expressed as

$\Pr[A_{n-1}=a$, a coalescence occurred at T_n, and resulting original line of descent lies in one of the a_{r-1} equivalence classes of size $r-1$ in $a]$

which may be further decomposed as

$$\sum_{i=1}^{a_{r-1}} ABC$$

where

$A=\Pr[A_{n-1}=a]$
$B=\Pr[\text{coalescence at } T_n \mid A_{n-1}=a]$
$C=\Pr[$resulting original line of descent lies in the class C_i of size $r-1$ in $a \mid A_{n-1}=a$ and a coalescence at $T_n]$.

Here $\{C_i \equiv C_i(a);\ 1 \leqslant i \leqslant a_{r-1}\}$ enumerates all the equivalence classes of size $r-1$ in a.

The random partition A_{n-1} involves only mutation and coalescence subsequent to T_n and by the Poisson nature of these events A_{n-1} is independent of the type of event which occurred at T_n. Consequently

$$B=\Pr[\text{coalescence at } T_n]$$
$$=\frac{n(n-1)}{2}\bigg/\frac{n(\theta+n-1)}{2}=\frac{n-1}{\theta+n-1}$$

in view of the competing Poissons.

Finally, the event C specifies $r-1$ of the $n-1$ lines of descent involving a. In view of the exchangeability present forcing lines of descent to be lost completely at random each of the $n-1$ original lines of descent has the same chance of representing the pair which coalesced at T_n. Thus

$$C=\frac{r-1}{n-1}.$$

This gives

$$\Pr[A_n=b,\ A_{n-1}=a]=\sum_{i=1}^{a_{r-1}}\Pr[A_{n-1}=a]\frac{n-1}{\theta+n-1}\,\frac{r-1}{n-1}$$
$$=\frac{(r-1)a_{r-1}}{\theta+n-1}\Pr[A_{n-1}=a].$$

The second case is

(2) $a_1=b_1-1$, the remaining $a_j=b_j$ $(j \leqslant n-1)$: This corresponds to a mutation at T_n causing one of the equivalence classes in S_n, having cardinality one, to be eliminated. The evaluation of $\Pr[A_n=b,\ A_{n-1}=a]$ proceeds as in case (1), though slightly simplified.

$$\Pr[A_n=b,\ A_{n-1}=a]=\Pr[A_{n-1}=a \text{ and a mutation occurred at } T_n]$$
$$=\Pr[A_{n-1}=a]\Pr[\text{mutation at } T_n \mid A_{n-1}=a]$$
$$=\Pr[A_{n-1}=a]\frac{\theta}{\theta+n-1}.$$

If we divide by $\Pr[A_{n-1}=a]$ we arrive at

$$\Pr[A_n=b\,|\,A_{n-1}=a]=\begin{cases}\dfrac{(r-1)a_{r-1}}{\theta+n-1} & \text{in (1),}\\[2ex] \dfrac{\theta}{\theta+n-1} & \text{in (2).}\end{cases}$$

The computation of the general one-step transition probabilities $\Pr[A_i=b\,|\,A_{i-1}=a]$, $1\leqslant i\leqslant n-1$ proceeds identically because in the evaluation of the joint probability $\Pr[A_i=b, A_{i-1}=a]$ in view of the Poisson nature of coalescence and mutation, the i original lines of descent may be looked upon as representing a sample of size i defining an i-coalescent upon which mutation is superimposed so that the previous arguments carry through with i replacing n, giving

$$\Pr[A_i=b\,|\,A_{i-1}=a]=\begin{cases}\dfrac{(r-1)a_{r-1}}{\theta+i-1} & \text{in (1)}\\[2ex] \dfrac{\theta}{\theta+i-1} & \text{in (2).}\end{cases} \tag{8.5}$$

These equations are identical with (7.4) the transition probabilities for the urn $\mathscr{U}$. Since A_1 and Π_1 both begin with the partition of one our proof is complete.

Corollary. *The jump chain $\{A_n, A_{n-1}, \ldots, A_1\}$ of the backward genealogical process $\mathscr{A}_t$ is equivalent to the reverse partition chain $\{\Pi_n, \Pi_{n-1}, \ldots, \Pi_1\}$ of the urn $\mathscr{U}$.*

Among the $\{T_n, T_{n-1}, \ldots, T_1\}$ will be k (random) times $T_{n_k}<T_{n_{k-1}}<\cdots<T_{n_1}\equiv T_1$ at which an original line of descent is created by a mutation. Denote these corresponding ancestors by $\mathscr{I}_k, \mathscr{I}_{k-1}, \ldots, \mathscr{I}_1$. By construction each individual in the sample must be descended without mutation from some $\mathscr{I}_i$ (because lines of descent are eliminated in chronological order beginning with the present). At time T_1 a single (the first) line of descent is created corresponding to $\mathscr{I}_1$. (We are running backwards through the coalescent, that is forward in genealogical time.) Label it one. At T_2 a second line is created. If it arose from the splitting of the first line then it (remains part of the first line and) keeps the label one. If it arose by mutation (of some other line, unspecified) then it must correspond to $\mathscr{I}_2$ and is so labelled two. In general, if the line of descent created at time T_i resulted from the splitting of an existing line, then it keeps the label of that line, else it is a mutation corresponding to some $\mathscr{I}_j$ and it is assigned the (previously unused) label j. All individuals which descend without mutation from $\mathscr{I}_j$ inherit label j. This results in a classification of the alleles by ages, in which the oldest is assigned 1, the second oldest 2, ... and so on. Since individuals are assigned the same age if and only if they trace back without mutation to a common ancestor the partition by ages gives, by definition, the same partition as the random equivalence relation $\mathscr{R}$ of Kingman.

Introduce marker variables $\{Y_1, Y_2, \ldots, Y_n\}$ where Y_i is the label assigned at time T_i.

Theorem 7. *The process $\{Y_1, Y_2, \ldots, Y_n\}$ is equal in distribution to the urn process $\{X_1, X_2, \ldots, X_n\}$.*

Proof. This is essentially a restatement of Theorem 6. The process $\{Y_1, Y_2, \ldots, Y_n\}$ bears the same relationship to $\{A_1, A_2, \ldots, A_n\}$ as $\{X_1, X_2, \ldots, X_n\}$ bears to $\{\Pi_1, \Pi_2, \ldots, \Pi_n\}$ in view of (8.5).

We are now in a position to explain the one-to-one correspondence between urn paths and equivalence relations described after (8.3). Each original line of descent acquires an age category based on Y_i. By construction there is no mutation along any line from the time it is created until the present. At time $t=0$ there are n original lines of descent corresponding to the n members of the initial sample. Each will have an age label Y_i (a label as an allele) and a time number (the subscript i on Y_i) indicating its chronological order of occurrence. Earlier we pointed out that numbering the lines of descent in the process L_t disturbs the simplicity of the structure. But suppose we number the members of the sample by their chronological number as just described. This will give a specific joint distribution to $(S_n, S_{n-1}, \ldots, S_1)$. In fact, if $S_n = \xi$ where ξ is an equivalence relation on $\mathscr{E}_n$ then automatically S_{n-1} is deterministically forced to be η where η is the equivalence relation induced on $\mathscr{E}_{n-1}$ from ξ by the deletion of n from $\{1, 2, \ldots, n\}$ because the last line created (numbered n) is lost first going back in real time. The same deterministic structure holds going through $S_{n-1}, S_{n-2}, \ldots, S_1$. But in the other direction, in view of (8.4)

$$\Pr[S_i = \xi \mid S_{i-1} = \eta] = \frac{\dfrac{\theta^k}{[\theta]^i} \displaystyle\prod_{i=1}^{k} (\lambda_i - 1)!}{\dfrac{\theta^m}{[\theta]^{i-1}} \displaystyle\prod_{j=1}^{m} (\mu_j - 1)!} \cdot P[S_{i-1} = \eta \mid S_i = \xi]$$

where $\{\lambda_i\colon 1 \leq i \leq k\}$ are the sizes of the equivalence classes of ξ and $\{\mu_j\colon 1 \leq j \leq m\}$ are the sizes of the equivalence classes in η. For any consistent pair (η, ξ) with the above numbering $\Pr[S_{i-1} = \eta \mid S_i = \xi] = 1$. Consequently

$$\Pr[S_i = \xi \mid S_{i-1} = \eta] = \begin{cases} \dfrac{r-1}{\theta + i - 1} & \text{if coalescence at } T_i \\ \dfrac{\theta}{\theta + i - 1} & \text{if mutation at } T_i \end{cases} \tag{8.6}$$

where $r-1$ is the size of the equivalence class of η involved in the coalescence.

Equation (8.6) is a restatment of (4.5) and (4.6). The former describes the probabilistic mechanism by which new lines of descent are created (either by the splitting of an existing line or by mutation), while the latter, describing the urn, gives the probabilities that each allele entering the sample is previously observed or novel. Watterson (1984, Sect. 3.6) and Donnelly and Tavaré (1986, Sect. 5) also give case (2) of (8.6) although only for the unconditioned probabilities Pr[mutation at T_i].

The thrust of Theorems 6 and 7 is that the urn process in the forward direction looks like the jump chain of the coalescent with mutation run backwards to the present from the birth time of the oldest allele in the sample. The labelling of alleles by the order in which they appear in the sample is probabilistically equivalent to labelling them by the order in which their ancestors appeared in genealogical time, and thus by their age. This justifies the methodology developed in Sect. 5 and we formalize this assertion as a corollary.

Corollary 1. *The distribution of allele numbers in the sample labelled by age is the same as the distribution of allele numbers labelled by the order in which they arise in the sample.*

The coalescent was introduced by Kingman as a robust approximation to the genealogy of a sample of individuals taken from a large population evolving according to one of a number of similar models. Suppose given a hypothetical infinite population to which the coalescent applies *exactly* and let O_i denote the proportion in the population of the ith oldest allele. Take a sample of size n and for fixed r let $E_{n,r}$ be the event that the oldest, second oldest, $\dots$, rth oldest alleles in the population are contained in the sample. By definition the oldest allele is always present in the population meaning that $\Pr[O_1>0]=1$ and thus $\Pr[\sum_{j=r+1}^{\infty} O_j<1]=1$. Therefore $(\sum_{j=r+1}^{\infty} O_j)^n \to 0$ with probability one as $n\to\infty$ and since $\Pr[E_{n,r}]=1-E[(\sum_{j=r+1}^{\infty} O_j)^n]$

$$\lim_{n\to\infty} \Pr[E_{n,r}]=1. \tag{8.7}$$

Denote by $N_i(n;S)$ the number of individuals in the sample of the ith oldest allele in the sample and let $N_i(n;P)$ be the number of individuals in the sample of the ith oldest allele in the population. As n increases, the oldest allele in the sample changes so that $N_i(n;S)$ fluctuates, but $N_i(n;P)$ of course grows monotonically. For each sample path eventually (by (8.7)) the first r oldest alleles in the population lie in the sample and $N_i(n;S)=N_i(n;P)$, $1\le i\le r$, $n\ge n_0$ (random).

Corollary 2. *The proportions* $(O_1, O_2, \dots)$ *in the population of the alleles ordered by decreasing age have the same representation* (2.1) *as the alleles relabelled by their order of observation in the sample.*

Proof. According to Theorem 7 and Corollary 1 the random vector $1/n(N_1(n;S),\dots,N_r(n;S))$ has the same joint distribution as $1/n(S_1(n),\dots,S_r(n))$ which by Theorem 1 converges a.s. to the residual allocation model described by (2.1). Therefore $1/n(N_1(n;P),\dots,N_r(n;P))$ has the same limiting distribution. But it converges to $(O_1,O_2,\dots,O_r)$ by the strong law of large numbers for exchangeable random variables (namely the sample from the population $\underset{\sim}{O}$).

This corollary implies Griffiths' result mentioned in Sect. 5. Our proof, based on the coalescent, should remain valid in the limit for any model in the domain of attraction of the coalescent. We will not pursue the appropriate limiting operations and weak convergence needed for a precise interpretation, since Corollary 1 is of primary importance. Incidentally, Corollary 1 does not follow from Corollary 2. Just because a population is described by (2.1) this does not imply that the first observation, for instance, will be from the category labelled one.

It is remarkable that Theorems 6 and 7 should hold. After all, the urn process $\{X_1,X_2,\dots,X_n\}$ is (by Theorem 3) a sample from a population, while $\{Y_1,Y_2,\dots,Y_n\}$ is the jumpchain of a stochastic process. The latter takes place in continuous (and in reverse) time while the former involves no temporal structure, the subscripts on the X merely marking the different alleles in the sample.

We close this section by touching base with recent work of Watterson (1984), Ethier and Griffiths (1987), and Donnelly and Tavaré (1986), all exploring the terrain of the coalescent. The last paper in particular derives joint distributions for age partitions and obtains Griffiths' representation for the age-class proportions. Our continuous time process $\mathcal{S}_t$ of random equivalence relations is novel and gives a genealogical interpretation of the urn by automatically tracing the ancestry of a sample back in time according to the ages (resulting in Theorem 6). The points of contact between $\mathcal{S}_t$ and the papers cited above deserve further study. It also seems appropriate to investigate the connection between the genealogical approach and reversibility, perhaps by way of an infinite dimensional diffusion whose sample paths are not delabelled order statistics (Ethier and Kurtz (1981, 1986)) but are refined enough to track the changing genealogy. We have some preliminary results in this direction.

9. Final remarks

There are numerous interwoven ideas in this paper which we briefly summarize.

This paper has analyzed the behaviour of a Pólya-like urn model which generates a Markov chain of partitions having the Ewens sampling formula as the marginal distributions and which imposes a labelling of alleles according to their order of occurrence in the sample. This is identical in distribution to that specified by the ages, a beautiful duality which we have shown ensues because the urn partition process is equivalent to the discrete skeleton representing the jumps in the time-reversed coalescent with mutation. This genealogical interpretation explains Griffiths' result that the age distribution of alleles in the population is the size-biased permutation of the Poisson–Dirichlet distribution.

The sequential nature of the urn often obviates the need to compute joint probabilities involving the Poisson–Dirichlet (or its size-biased sibling) and provides a simple basis for computing age and copy number distributions of alleles, both in a sample and the population, as manifested by the numerous examples above. In the spirit of this technique, we have built, on the urn, a direct proof that a random sample from a Poisson–Dirichlet population has the Ewens sampling structure. Previously this has been verified by passage to the limit from a Dirichlet distribution because of the difficulties of integration with respect to the Poisson–Dirichlet. Furthermore this approach illuminates Ewens' original prescription of the appearance of alleles in the sample.

The Poisson–Dirichlet has a representation as an infinite residual allocation model when described in size-biased form. We have formulated a concept of residual allocation model appropriate for discrete populations and have shown that the size-biased version of Ewens' partition is such a model which is additionally invariant under size-biased permutation (analogously to the Poisson–Dirichlet) and is equivalent to partition by ages of alleles. The representation provides a new method for efficient simulation from a Poisson–Dirichlet population as well as from any other residual allocation model.

Acknowledgments. It is a pleasure to thank Warren Ewens for kindly giving me his comments on an earlier version of this paper, Simon Tavaré for a number of useful discussions, Geoff Watterson for providing historical information, and a referee for a splendid job in reviewing the manuscript. His careful reading uncovered a number of errors and his insistence on detail caused me to rethink some proofs, resulting in a vastly improved exposition, for which I am most grateful.

References

Blackwell, D., MacQueen, J. B.: Ferguson distributions via Pólya urn schemes. Ann. Statist. **1**, 353-355 (1973)

Connor, R. J., Mosimann, J. E.: Concepts of independence for proportions with a generalization of the Dirichlet distribution. J. Am. Statist. Assoc. **64**, 194-206 (1969)

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

Engen, S.: A note on the geometric series as a species frequency model. Biometrica **62**, 694-699 (1975)

Ethier, S. N., Griffiths, R. C.: The infinitely-many-sites model as a measure-valued diffusion. Ann. Probab., to appear (1987)

Ethier, S. N., Kurtz, T. G.: The infintely-many-neutral-alleles diffusion model. Adv. Appl. Probab. **13**, 429-452 (1981)

Ethier, S. N., Kurtz, T. G.: Markov processes; characterization and convergence. New York: Wiley 1986

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

Ewens, W. J.: Testing for increased mutation rate for neutral alleles. Theor. Popul. Biol. **4**, 251-258 (1973)

Ewens, W. J.: Mathematical population genetics. New York Heidelberg Berlin: Springer 1979

Fuerst, P. A., Chakraborty, R., Nei, M.: Statistical studies on protein polymorphism in natural populations. I. Distribution of single locus heterozygosity. Genetics **86**, 455-483 (1977)

Good, I. J.: The estimation of probabilities. Cambridge: MIT Press 1965

Griffiths, R. C.: Lines of descent in the diffusion approximation of neutral Wright-Fisher models. Theor. Popul. Biol. **17**, 37-50 (1980)

Griffiths, R. C., Li, W.-H.: Simulating allele frequencies in a population and the genetic differentiation of populations under mutation pressure. Theor. Popul. Biol. **23**, 19-33 (1983)

Hill, B. M.: Posterior moments of the number of species in a finite population and the posterior probability of finding a new species. J. Am. Statist. Assoc. **74**, 668-673 (1979)

Hoppe, F. M.: Pólya-like urns and the Ewens sampling formula. J. Math. Biol. **20**, 91-99 (1984)

Hoppe, F. M.: Size-biased filtering of Poisson-Dirichlet samples with an application to partition structures in genetics. J. Appl. Probab. **23**, 1008-1012 (1986)

Karlin, S., McGregor, J.: The number of mutant forms maintained in a population. Proc. Fifth. Berk. Symp. Math. Stat. and Prob. II, 415-438 (1967)

Karlin, S., McGregor, J.: Addendum to a paper of W. Ewens. Theor. Popul. Biol. **3**, 113-116 (1972)

Kelly, F. P.: Exact results for the Moran neutral allele model. Adv. Appl. Probab. **9**, 197-201 (1977)

Kingman, J. F. C.: Random discrete distributions. J. Roy. Statist. Soc. B. **37**, 1-22 (1975)

Kingman, J. F. C.: The population structure associated with Ewens' sampling formula. Theor. Popul Biol. **11**, 274-283 (1977)

Kingman, J. F. C.: Random partitions in population genetics. Proc. R. Soc. Lond. A. **361**, 1-20 (1978)

Kingman, J. F. C.: The mathematics of genetic diversity. CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. **34**. Philadelphia, PA S.I.A.M. 1980

Kingman, J. F. C.: On the genealogy of large populations. J. Appl. Prob. **19A**, 27-43 (1982a)

Kingman, J. F. C.: The coalescent. Stoch. Proc. Applic. **13**, 235-248 (1982b)

Kingman, J. F. C.: Exchangeability and the evolution of large populations. In: Koch, G., Spizzichino, F. (eds.) Exchangeability in probability and statistics. Amsterdam: North Holland 1982c

McCloskey, J.W.: A model for the distribution of individuals by species in an environment. Ph.D. thesis, Michigan State University (1965)

Patil, G. P., Taillie, C.: Diversity as a concept and its implications for random communities. Bull. Int. Stat. Inst. **XLVII**, 497-515 (1977)

Saunders, I. W., Tavaré, S., Watterson, G. A.: On the genealogy of nested subsamples from a haploid population. Adv. Appl. Prob. **16**, 471-491 (1984)

Tavaré, S.: Line-of-descent and genealogical processes, and their applications in population genetics models. Theor. Popul. Biol. **26**, 119-164 (1984)

Trajstman, A. C.: On a conjecture of G. A. Watterson. Adv. Appl. Prob. **6**, 489-493 (1974)

Watterson, G. A.: Models for logarithmic species abundance distributions. Theor. Popul. Biol. **6**, 217-250 (1974)

Watterson, G. A.: The sampling theory of selectively neutral alleles. Adv. Appl. Prob. **6**, 463-488 (1974)

Watterson, G. A.: The stationary distribution of the infinitely-many neutral alleles diffusion model. J. Appl. Probab. **13**, 639-651 (1976)

Watterson, G. A.: Reversibility and the age of an allele. I. Moran's infinitely-many neutral alleles model. Theor. Popul. Biol. **10**, 239-253 (1976)

Watterson, G. A.: Lines of descent and the coalescent. Theor. Popul. Biol. **26**, 72-92

Watterson, G. A.: Estimating the divergence time of two species, to appear (1985)

Watterson , G. A., Guess, H. A.: Is the most frequent allele the oldest? Theor. Popul. Biol. **11**, 141-160 (1977)

Wilks, S. S.: Mathematical statistics. New York: Wiley 1962

Received September 2, 1985/Revised July 16 and November 11, 1986

Note added in proof. P. Donnelly (Theor. Popul. Biol. **30**, 271–288) has, meanwhile, also discussed the above urn model and its relation to the ages of alleles.