Some Developments of the Blackwell-MacQueen Urn Scheme

Jim Pitman
University of California

August 29, 1996

Abstract

The Blackwell-MacQueen description of sampling from a Dirichlet random distribution on an abstract space is reviewed, and extended to a general family of random discrete distributions. Results are obtained by application of Kingman's theory of partition structures.

1 Introduction

Blackwell and MacQueen [10] described the construction of a Dirichlet prior distribution by a generalization of Pólya's urn scheme. While the notion of a random discrete probability measure governed by a Dirichlet distribution was first developed in the setting of Bayesian statistics [30, 26, 27, 28], this idea has applications in other fields. The distribution of the ranked masses of atoms in a Dirichlet distribution, called the *Poisson-Dirichlet* (PD) *distribution* [45], appears as an asymptotic distribution in number theory [14, 8, 67, 16], combinatorics [65, 68, 69, 34], and population genetics [70, 24]. Though the finite dimensional distributions of the PD distribution are difficult to describe explicitly, there are some remarkably simple formulae involving this distribution, most notably the Ewens sampling formula [23, 25]. Antoniak [3] derived the Ewens sampling formula from the Blackwell-MacQueen description of sampling from a Dirichlet prior. Hoppe [35, 37] used the urn scheme to derive the simple form of the *size-biased random permutation* of the PD distribution, which Ewens [24] termed the GEM distribution, after Griffiths, Engen and McCloskey, who contributed to its development and application in the fields of genetics and ecology. Dirichlet random measures and the PD and GEM distributions appear also as the stationary distributions of measure-valued diffusions derived from population genetics models [19, 20, 21, 22].

Section 2 of this paper reviews some basic results involving Dirichlet distributions and the PD and GEM distributions. Section 3 shows how many of these results extend to a more general Bayesian model for sampling from a random discrete distribution. This involves Kingman's theory of partition structures [46] as developed in [1, 55]. The general model is illustrated by a two-parameter model for species sampling, first proposed by Engen [18], which is defined here following [55] by a variation of the Blackwell-MacQueen urn scheme. This two-parameter model is a natural extrapolation of a basic model for species sampling proposed by R.A. Fisher in 1943 [29]. The family of random discrete distributions associated with this two-parameter model, which can be characterized in a number of ways [53, 58, 71, 43], turns out to include both Dirichlet distributions and distributions derived from the lengths of excursions of Brownian motion and Bessel processes [53, 59, 54, 60].

Finally, Section 4 indicates briefly how the Blackwell-MacQueen urn scheme and its generalizations described in Section 3 can be interpreted in terms of random permutations.

2 Dirichlet Distributions

2.1 Preliminaries

For $\mu > 0$ let gamma(μ) denote the gamma distribution on $(0, \infty)$ with mean μ, whose density at x is $\Gamma(\mu)^{-1}x^{\mu-1}e^{-x}$, and define gamma(0) to be the distribution degenerate at 0. Recall that if $\Gamma_1, \dots, \Gamma_k$ are independent and Γ_i has gamma(μ_i) distribution, then the distribution of the random vector $(Y_1, \dots, Y_k)$, where $Y_i = \Gamma_i / \Sigma_{j=1}^k \Gamma_j$, is called *Dirichlet with parameter* $(\mu_1, \dots, \mu_k)$, denoted by $\text{DIRICHLET}(\mu_1, \dots, \mu_k)$. Note that the ith component of a random vector with $\text{DIRICHLET}(\mu_1, \dots, \mu_k)$ distribution has a beta(μ_i $, \sum_{j \neq i} \mu_j$) distribution on $[0, 1]$, where for $a > 0$ and $b > 0$ the beta(a, b) distribution has density $\Gamma(a+b)\Gamma(a)^{-1}\Gamma(b)^{-1}x^{a-1}(1-x)^{b-1}$ at $x \in (0, 1)$. Let $(S, \mathcal{S})$ be an abstract measurable space. To avoid measure theoretic pathologies, it is assumed throughout that

$$\text{the diagonal } \{(x, y) : x = y\} \text{ is a product measurable subset of } S \times S \quad (1)$$

Call F a *random distribution* on S if F is a collection of random variables

$$F = (F(B), B \in \mathcal{S}) = (F(B, \omega), B \in \mathcal{S}, \omega \in \Omega) \quad (2)$$

defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, such that for each $\omega \in \Omega$, the function $B \to F(B, \omega)$ is a probability measure on $(S, \mathcal{S})$. Call a sequence of random variables (X_n) a *sample from* F if conditionally given F the X_n are independent and identically distributed according to F, abbreviated i.i.d.(F).

Let μ be a positive measure on $(S, \mathcal{S})$ with $0 < \mu(S) < \infty$. Say F has DIRICHLET(μ) distribution if F is a random distribution on S such that for every measurable partition $B_1, \ldots, B_k$ of S, the random vector $(F(B_1), \ldots, F(B_k))$ has DIRICHLET$(\mu(B_1), \ldots, \mu(B_k))$ distribution. Ferguson [27] established the existence of such a random distribution, and proved the following extension of the well known updating rule for sampling from a Dirichlet prior on a finite set. For $x \in S$ let $\delta(x)$ denote the distribution of a unit mass at x. So $\delta(x, B) = 1(x \in B), B \in \mathcal{S}$.

Theorem 1 [27]. *If (X_n) is a sample from F with* DIRICHLET(μ) *distribution, then the conditional distribution of F given $X_1, \ldots, X_n$ is* DIRICHLET(μ_n) *where μ_n is the random measure*

$$\mu_n = \mu + \sum_{i=1}^{n} \delta(X_i) \tag{3}$$

If (X_n) is a sample from a DIRICHLET(μ) prior F, then the unconditional distribution of each X_n is $\mu/||\mu||$ where $||\mu|| = \mu(S)$ is the total mass of μ. So Theorem 1 implies the *Blackwell-MacQueen prediction rule*[10]:

$$P(X_{n+1} \in \cdot \,|\, X_1, \ldots, X_n) = \mu_n(\cdot)/||\mu_n|| \tag{4}$$

Blackwell and MacQueen reversed this derivation of (4) to establish the following result:

Theorem 2 [10] *Let (X_n) be a sequence of random variables constructed so that X_1 has distribution $\mu/||\mu||$, and (4) holds for μ_n as in (3). Let $F_n = \mu_n/||\mu_n||$. Then*

(i) *F_n converges a.s. as $n \to \infty$ to a random discrete distribution F;*

(ii) *F has Dirichlet (μ) distribution;*

(iii) *$X_1, X_2, \ldots$ is a sample from F.*

Blackwell and MacQueen assumed that $(S, \mathcal{S})$ is a Polish space with Borel σ-field, and proved weak convergence of F_n to F almost surely. But it will be seen in Section 3 that Theorem 2 holds under the weaker regularity condition (1), with the convergence in (i) meaning convergence in total variation almost surely. See also Blackwell [9] and Berk-Savage [7] regarding the discreteness of Dirichlet distributions.

2.2 Ranked and size-biased frequencies

For a measure μ on S with $0 < \mu(S) < \infty$ let $\theta = \mu(S)$ and $\nu = \mu/\theta$. So $\theta > 0$, ν is a probability distribution on S, and $\mu = \theta\nu$.

Theorem 3 [27] *Let* $\Gamma_{(1)} > \Gamma_{(2)} > \cdots$ *be the points of a Poisson random measure on* $(0, \infty)$ *with mean measure* $\theta x^{-1} e^{-x} dx$. *Put*

$$P_i = \Gamma_{(i)}/\Sigma \text{ where } \Sigma = \sum_i \Gamma_{(i)} \tag{5}$$

and define

$$F = \sum_{i=1}^{\infty} P_i \, \delta(\hat{X}_i) \tag{6}$$

where the $\hat{X}_i$ *are i.i.d.* (ν), *independent also of the* $\Gamma_{(i)}$. *Then* F *has* DIRICHLET$(\theta\nu)$ *distribution, independently of* Σ *which has gamma*(θ) *distribution.*

Definition 4 [45] The distribution of the sequence (P_i) defined by (5), with $P_1 > P_2 > \cdots > 0$ and $\sum_i P_i = 1$ almost surely, is called the *Poisson Dirichlet distribution with parameter* θ, abbreviated PD(θ).

Explicit formulae for the finite dimensional distributions of PD(θ) are known but rather complicated [70, 38, 52]. The construction (5) of (P_i) with PD(θ) distribution is related to a derivation of Dirichlet distributions from Fisher's [29] model for species sampling, which is now described. Suppose there are m distinct species in a population and that individuals of the ith species are trapped according to a homogeneous Poisson process with rate Γ_i where $\Gamma_1, \ldots, \Gamma_m$ are i.i.d. with gamma(κ) distribution for some $\kappa > 0$. Call Γ_i the *abundance* of the ith species. Let $\Gamma_{(1)} > \cdots > \Gamma_{(m)}$ denote the order statistics of these abundances. As noted by Fisher and others [2, 45, 70, 18, 17], various features of this model have simple limits as $m \to \infty$ and $\kappa \to 0$ with $m\kappa = \theta$ held fixed. From a modern perspective, these limits are features of a limiting model with ranked abundances $\Gamma_{(1)} > \Gamma_{(2)} > \cdots$ defined by ranking the points of a Poisson process as in Theorem 3. In this limit model, introduced by McCloskey [50], the sampling process records arrivals from an infinite collection of independent Poisson processes with random rates $\Gamma_{(i)}$. Suppose the jth species to be observed in such a sampling process is the species whose abundance is $\Gamma_{(\pi_j)}$ and let $\tilde{P}_j = P_{\pi_j}$ be the corresponding relative frequency of this species. Elementary properties of Poisson processes imply that $(\tilde{P}_j)$ is a *size-biased permutation* of (P_i). That is to say, $\tilde{P}_j = P_{\pi_j}$ where for all finite sequences $(i_j, 1 \leq j \leq k)$ of distinct positive integers, the conditional probability of the event $(\pi_j = i_j$ for all $1 \leq j \leq k)$ given $(P_1, P_2, \ldots)$ is

$$P_{i_1} \frac{P_{i_2}}{1 - P_{i_1}} \cdots \frac{P_{i_k}}{1 - P_{i_1} - \ldots - P_{i_{k-1}}} \tag{7}$$

Theorem 5 [50] *Let* $(\tilde{P}_j)$ *be a size-biased permutation of a sequence of random variables* $P_1 > P_2 > \cdots > 0$ *with* $\sum_i P_i = 1$. *Then*

$$\tilde{P}_j = \left[\prod_{i=1}^{j-1}(1-\tilde{W}_i)\right]\tilde{W}_j \tag{8}$$

for a sequence of i.i.d. random variables $(\tilde{W}_j)$ *iff* (P_i) *has* PD (θ) *distribution for some* $\theta > 0$. *The common distribution of the* $\tilde{W}_j$ *is then beta*$(1,\theta)$

McCloskey's thesis [50] is unpublished, but a proof of Theorem 5 can be found in [53]. The model (8) for generating a random discrete distribution $(\tilde{P}_j)$ from independent factors $\tilde{W}_i$ has been studied by many authors [33, 30, 12, 26, 15, 51]. See also [36, 17, 58] for further study of size-biased permutations and their applications.

Definition 6 [24] Say that a sequence of random variables $(\tilde{P}_j)$ has GEM (θ) distribution iff (8) holds for a sequence $(\tilde{W}_j)$ of i.i.d. beta$(1,\theta)$ variables.

McCloskey's theorem has the following corollary:

Corollary 7 *Let* (P_i) *with* $P_1 \geq P_2 \geq \cdots$ *be defined by ranking a sequence* $(\tilde{P}_j)$ *with* GEM (θ) *distribution. Then*

(i) (P_i) *has* PD (θ) *distribution, and*

(ii) $(\tilde{P}_j)$ *is a size-biased permutation of* (P_i).

Proof. Part (i) is an obvious consequence of Theorem 5. To see why (ii) is true, write simply $\tilde{\boldsymbol{P}}$ for $(\tilde{P}_j)$ and $\boldsymbol{P}$ for (P_i). Then $\boldsymbol{P} = r(\tilde{\boldsymbol{P}})$ where the ranking function r is a product measurable function on sequence space. Because the values P_i are a.s. distinct, it is clear that $\tilde{P}_j = P_{\pi_j}$ for an a.s. uniquely defined sequence of random variables (π_j), and hence that $\tilde{\boldsymbol{P}}$ is a size-biased permutation of $\boldsymbol{P}$ iff $\tilde{\boldsymbol{P}}$ has a particular conditional distribution given $\boldsymbol{P}$. But a conditional distribution of $\tilde{\boldsymbol{P}}$ given $r(\tilde{\boldsymbol{P}})$ that serves for some sequence $\tilde{\boldsymbol{P}}$ with a prescribed distribution must work for every sequence $\tilde{\boldsymbol{P}}$ with that distribution. $\Box$

Combining Theorems 3 and 5 yields the following result:

Corollary 8 [64, 63]. *Let* F *be defined by*

$$F = \sum_{j=1}^{\infty} \tilde{P}_j\, \delta(\tilde{X}_j) \tag{9}$$

for two sequences of random variables $(\tilde{P}_j)$ *and* $\tilde{X}_j$ *such that* $(\tilde{P}_j)$ *has the* GEM (θ) *distribution* (8) *and the* $\tilde{X}_j$ *are i.i.d* (ν), *independent of* $(\tilde{P}_j)$. *Then* F *has* DIRICHLET$(\theta\nu)$ *distribution.*

This construction provides both a simple way to simulate F with DIRICHLET(μ) distribution, and an approach to computation of the distribution of functionals of F such as

$$\int_S g\,dF = \sum_{j=1}^{\infty} \tilde{P}_j g(\tilde{X}_j) \tag{10}$$

for a measurable function g on S. For further developments see [61, 62, 13, 44].

Call μ *diffuse* if $\mu(\{x\}) = 0$ for all $x \in S$. Then the $\hat{X}_i$ in (6) are a.s. distinct, and the relation between the two constructions (9) and (6) of a DIRICHLET(μ) distributed F can be clarified as follows:

Corollary 9 *Suppose F has* DIRICHLET($\theta\nu$) *distribution, for $\theta > 0$ and a diffuse probability distribution ν on S. Let P_i denote the magnitude of the ith largest atom of F, and let $\hat{X}_i$ be the location of this atom in the space S. Let $\tilde{X}_j$ denote the jth distinct value observed in a sample (X_n) from F and let $\tilde{P}_j = F(\{\tilde{X}_j\})$, the size of the atom of F at $\tilde{X}_j$. Then almost surely*

$$F = \sum_{i=1}^{\infty} P_i\,\delta(\hat{X}_i) = \sum_{j=1}^{\infty} \tilde{P}_j\,\delta(\tilde{X}_j) \tag{11}$$

where

(i) $(P_1, P_2, \ldots)$ *has* PD(θ) *distribution;*
(ii) *the $\hat{X}_i$ are i.i.d (ν), independently of (P_i);*
(iii) $(\tilde{P}_j)$ *is a size-biased permutation of (P_i);*
(iv) $(\tilde{P}_1, \tilde{P}_2, \ldots)$ *has* GEM(θ) *distribution;*
(v) *the $\tilde{X}_j$ are i.i.d (ν), independently of $(\tilde{P}_j)$.*

Corollary 10 *Suppose that F with* DIRICHLET($\theta\nu$) *distribution, for $\theta > 0$ and a diffuse probability measure ν, is constructed via (9) for $(\tilde{P}_j)$ and $\tilde{X}_j$ as in Corollary 8, then expressed in terms of its ranked atoms (P_i) and their locations $\hat{X}_i$, so that formula* (11) *holds by construction. Then the joint distribution of the four sequences $(\tilde{P}_j)$, $(\tilde{X}_j)$, (P_i), and $(\hat{X}_i)$ is as described in parts* (i)-(v) *of Corollary 9. In particular, $(\tilde{P}_j)$ is a size-biased random permutation of (P_i), say $\tilde{P}_j = P_{\pi_j}$, and $\tilde{X}_j = X_{\pi_j}$.*

Proof. The subtlety here is that $\tilde{X}_j$ is not defined as the jth distinct value to appear in some random sample (X_n) from F, as in Corollary 9. It is only asserted that the joint distribution of $(\tilde{X}_j)$ and F is the same as if $\tilde{X}_j$ were the jth distinct value to appear in a sample from F, and this is a consequence of Corollary 7. □

3 Species Sampling Models

As noted by Ferguson [27], the fact that Dirichlet distributions are a.s. discrete has some disadvantages from the standpoint of conventional non-parametric problems of Bayesian inference, where it may be preferable to have an a.s. continuous prior with an explicit updating rule [49]. But there is one setting where discreteness of a prior distribution is very much a positive feature. This is in problems of *species sampling*, as studied in ecology [29, 18], population genetics [24], and other settings [66, 11]. Suppose that a random sample $X_1, X_2, \ldots$ is drawn from a large population of individuals of various *species*, and X_i represents the species of the ith individual sampled. The space S in this setting should be thought of as an arbitrary set of tags, or a spectrum of colors, used to label various species. Following Aldous [1], it can be assumed that S is the unit interval and that the jth distinct species to appear in the sample is deliberately assigned a tag $\tilde{X}_j$ in $[0,1]$, where the $\tilde{X}_j$ are i.i.d. uniform $[0,1]$ variables generated by some additional randomization. Since the $\tilde{X}_j$ are a.s. distinct, different species are coded by different tags almost surely. This device of random tagging transforms an unconventional problem, where the observation at stage n is a random partition of n individuals into various species, into a conventional one where what is observed at stage n is a sequence of tags $(X_1, \ldots, X_n)$.

More formally, let (X_n) be a sequence of random variables with values in $(S, \mathcal{S})$, defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $M_1 := 1$ and for $j > 1$ let

$$M_j := \inf\{n : n > M_{j-1}, X_n \notin \{X_1, \ldots, X_{n-1}\}\} \tag{12}$$

with the convention $\inf \emptyset := \infty$. On the event $M_j < \infty$ define $\tilde{X}_j := X_{M_j}$. In the language of species sampling, X_n represents the *species* of the nth *individual* in some process of sampling of individuals from a population, and $\tilde{X}_j$ is the *jth species to appear*. For $j = 1, 2, \ldots$ let N_{jn} be the number of times that the jth species $\tilde{X}_j$ appears in the sample $X_1, \ldots, X_n$:

$$N_{jn} := \sum_{m=1}^{n} 1(X_m = \tilde{X}_j, M_j < \infty) \tag{13}$$

Let $K_n := \max\{j : N_{jn} > 0\}$, the number of different species to appear in the first n observations.

Call a rule specifying the distribution of X_1 and the conditional distribution of X_{n+1} given $X_1, \ldots, X_n$ for each $n = 1, 2, \ldots$ a *prediction rule*. It will be assumed throughout that X_1 has some fixed probability distribution ν on S which is *diffuse*, that is $\nu(\{x\}) = 0$ for all $x \in S$. The symbol ν is a mnemonic for the distribution of a *new* species. To be definite, it can be supposed that $S = [0,1]$ and ν is uniform. For $\theta > 0$ the Blackwell-MacQueen prediction rule (4) for random sampling from a Dirichlet $(\theta\nu)$

random measure F can be written using the above notation as follows: for $1 \leq k \leq n$

$$\mathbb{P}(X_{n+1} \in \cdot \mid X_1, \ldots, X_n, K_n = k) = \sum_{j=1}^{k} \frac{N_{jn}}{n+\theta} 1(\tilde{X}_j \in \cdot) + \frac{\theta}{n+\theta} \nu(\cdot) \quad (14)$$

As a generalization, consider (X_n) subject to a prediction rule of the form

$$\mathbb{P}(X_1 \in \cdot\,) = \nu(\cdot) \quad (15)$$

$$\mathbb{P}(X_{n+1} \in \cdot \mid X_1, \ldots, X_n, K_n = k) = \sum_{j=1}^{k} p_j(\boldsymbol{N}_n) 1(\tilde{X}_j \in \cdot) + p_{k+1}(\boldsymbol{N}_n)\nu(\cdot) \quad (16)$$

where $\boldsymbol{N}_n := (N_{1n}, N_{2n}, \ldots)$ is the vector of counts of various species observed in the sample $(X_1, \ldots, X_n)$. Here the range of the random vectors $\boldsymbol{N}_n$ is identified in an obvious way with the countable set $\mathbb{N}^* := \bigcup_{k=1}^{\infty} \mathbb{N}^k$, the set of finite sequences of positive integers, and $(p_j, j = 1, 2, \ldots)$ is a sequence of *prediction probability functions* defined on $\mathbb{N}^*$. The functions p_j should be understood as follows. Given that after n observations the vector of counts of various species in order of appearance is $\boldsymbol{N}_n = \boldsymbol{n}$ say, where $\boldsymbol{n} = (n_1, \ldots, n_k) \in \mathbb{N}^*$ with $\sum_i n_i = n$, the next observation is the jth species already observed with probability $p_j(\boldsymbol{n})$ for $1 \leq j \leq k$, and a new species with probability $p_{k+1}(\boldsymbol{n})$. Here $k = k(\boldsymbol{n})$ is the number of non-zero components of $\boldsymbol{n}$, so the random number K_n of different species to appear in the first n observations is $K_n = k(\boldsymbol{N}_n)$, and

$$p_j(\boldsymbol{n}) = \mathbb{P}(X_{n+1} = \tilde{X}_j \mid \boldsymbol{N}_n = \boldsymbol{n}) \qquad (1 \leq j \leq k(\boldsymbol{n}) + 1) \quad (17)$$

It is clear that any sequence of functions p_j such that

$$p_j(\boldsymbol{n}) \geq 0, \qquad \sum_{j=1}^{k(\boldsymbol{n})+1} p_j(\boldsymbol{n}) = 1 \qquad (\boldsymbol{n} \in \mathbb{N}^*) \quad (18)$$

determines the distribution of a sequence of random variables (X_n) via the prediction rule (16). For example, the Blackwell-MacQueen rule (14) is the special case of (16) with

$$p_j(n_1, \ldots, n_k) = \frac{n_j}{n+\theta} 1(1 \leq j \leq k) + \frac{\theta}{n+\theta} 1(j = k+1) \quad (19)$$

where $n = \sum_{i=1}^{k} n_i$. The following proposition is an easy consequence of Kingman's theory of exchangeable random partitions as developed in [1, 55]:

Proposition 11 *Suppose (X_n) is an exchangeable sequence of random variables subject to a prediction rule of the form* (15)–(16). *Let F_n denote the conditional distribution of X_{n+1} given $X_1, \ldots, X_n$, as displayed in* (16).

(i) *F_n converges in total variation norm almost surely as $n \to \infty$ to the random distribution*

$$F := \sum_j \tilde{P}_j \delta(\tilde{X}_j) + (1 - \sum_j \tilde{P}_j)\nu \tag{20}$$

where $\tilde{P}_j$ is the frequency of the jth species to appear, that is

$$\tilde{P}_j := \lim_{n\to\infty} \frac{N_{jn}}{n} \text{ almost surely} \tag{21}$$

(ii) *the $\tilde{X}_j$ are i.i.d.(ν) independent of the $\tilde{P}_j$.*
(iii) $(X_1, X_2, \ldots)$ *is a sample from F.*

To be careful about the meaning of part (ii), note that the number K_∞ of distinct values in the infinite sequence $(X_1, X_2, \ldots)$ is almost surely equal to $\inf\{k : \tilde{P}_1 + \cdots + \tilde{P}_k = 1\}$. The meaning of (ii) is that conditionally given $(\tilde{P}_1, \tilde{P}_2, \ldots)$ with $K_\infty = k$ the $\tilde{X}_j$ are i.i.d.(ν) for $1 \leq j < k+1$. Such amplifications are required to interpret similar independence statements made below.

When compared to Theorem 2 for the Blackwell-MacQueen prediction rule, Proposition 11 is deficient in two respects. Firstly, Proposition 11 makes the assumption that (X_n) is exchangeable, which is part of the conclusion of the Blackwell-MacQueen theorem. Secondly, Proposition 11 provides no explicit description of the distribution of F. These deficiencies are remedied to some extent by the following discussion.

Definition 12 Call (X_n) a *species sampling sequence* if (X_n) is an exchangeable sequence subject to a prediction rule of the form (15)–(16) for a diffuse distribution ν, as supposed in Proposition 11.

As a variation of Proposition 11, it is easily seen that (X_n) is a species sampling sequence iff (X_n) is a sample from a random distribution F of the form

$$F = \sum_i P_i \delta(\hat{X}_i) + (1 - \sum_i P_i)\nu \tag{22}$$

for some sequence of random variables (P_i) such that

$$P_i \geq 0 \text{ and } \textstyle\sum_i P_i \leq 1 \text{ a.s.} \tag{23}$$

and some sequence $(\hat{X}_i)$ that is i.i.d.(ν) independent of (P_i).

This set-up, with a random distribution F of the form (22), and a sample (X_n) from F, will be called a *species sampling model.* Interpret P_i as the relative frequency of the ith species in some listing of species present in a population, and $\hat{X}_i$ as the tag assigned to that species. The random distribution F has an atom of magnitude P_i at $\hat{X}_i$ for each i such that $P_i > 0$, and the rest of its mass distributed proportionally to ν. Call the model *proper* if $\sum_i P_i = 1$ a.s.. That is to say, F is almost surely discrete. Then

$$F = \sum_i P_i \delta(\hat{X}_i) = \sum_j \tilde{P}_j \delta(\tilde{X}_j) \tag{24}$$

where $\tilde{X}_j$ and $\tilde{P}_j$ as in (20) are defined in terms of a sample (X_n) from F. Provided the sample (X_n) is conditionally i.i.d.(F) given both F and (P_i), as is the case if (P_i) is decreasing, the sequence $(\tilde{P}_j)$ is a size-biased permutation of (P_i). So (24) generalizes the representations (11) for F with DIRICHLET$(\theta\nu)$ distribution. It is easily verified that a species sampling model is proper iff

$$\mathbb{P}(\tilde{P}_1 > 0) = 1 \tag{25}$$

Alternatively, by application of the strong law of large numbers after conditioning on all the (P_i), the model is proper iff

$$\mathbb{P}(\lim_n K_n/n = 0) = 1 \tag{26}$$

3.1 The finite-dimensional distributions of F

Assuming that $S = [0,1]$ and ν is uniform, the random distribution F on $[0,1]$ constructed via (22) is determined by its cumulative distribution function $(F[0,t], 0 \le t \le 1)$, which is a random process with exchangeable increments. Moreover this construction yields the most general possible distribution for a random distribution function on $[0,1]$ with exchangeable increments [41]. For an abstract $(S, \mathcal{S}, \nu)$ the construction (22) yields the most general possible possible random distribution F whose finite dimensional distributions are described as follows: for every measurable partition $B_1, \ldots, B_k$ of S with $\nu(B_i) = t_i$ say, $(F(B_1), F(B_2), \ldots, F(B_k))$ has the same joint distribution as

$$(G(t_1), G(t_1 + t_2) - G(t_1), \ldots, 1 - G(t_1 + \cdots + t_{k-1})) \tag{27}$$

where $(G(t), 0 \le t \le 1)$ is the random distribution function on $[0,1]$ with exchangeable increments obtained from the model with ν uniform on $[0,1]$ and the same random frequencies. Such random measures were studied by Kallenberg [40, 42].

A large class of models, including the Dirichlet and its two-parameter extension described below, is obtained by supposing that $G(t) = Y(t)/Y(1)$ for

Y an increasing process with independent increments, and either conditioning on $Y(1)$ or allowing a change of measure by a density which is function of $Y(1)$. See [45, 53, 52, 56, 60, 54] for further study of these models.

3.2 The exchangeable partition probability function

Let $[n] = \{1, \ldots, n\}$, and for a finite set A let $\#A$ denote the number of elements of A. If (X_n) is exchangeable, then for each partition of $[n]$ into k non-empty subsets $A_1, \ldots, A_k$, where it is assumed that the A_i are in *order of appearance*, that is $1 \in A_1$, and for each $2 \leq j \leq k$ the first element of $[n] - (A_1 \cup \ldots \cup A_{j-1})$ belongs to A_j,

$$\mathbb{P}\left(\bigcap_{j=1}^{k}(X_\ell = \tilde{X}_j \text{ for all } \ell \in A_j)\right) = p(\#A_1, \ldots, \#A_k) \tag{28}$$

for some *symmetric* function p of k-tuples of non-negative integers with sum n. Allowing n to vary defines a function $p : \mathbb{N}^* \to [0,1]$, where $\mathbb{N}^* = \bigcup_{k=1}^{\infty} \mathbb{N}^k$ as before. This symmetric function p determines the distribution of the random partition of $\mathbb{N}$ whose classes are the equivalence classes for random equivalence relation defined by $i \sim j$ iff $X_i = X_j$. See [55] for further discussion. Call p the *exchangeable partition probability function* (EPPF) derived from the exchangeable sequence (X_n). As before, identify $\boldsymbol{n} = (n_1, \ldots, n_k) \in \mathbb{N}^*$ with the infinite sequence $(n_1, \ldots, n_k, 0, 0 \ldots)$ obtained by padding with zeros, and view the number k of non-zero components of $\boldsymbol{n}$ as a function $k = k(\boldsymbol{n})$ With this identification, for each sequence $\boldsymbol{n} \in \mathbb{N}^*$ and each $1 \leq j \leq k(\boldsymbol{n}) + 1$, a sequence $\boldsymbol{n}^{j+} \in \mathbb{N}^*$ is defined by incrementing n_j by 1. From the definition (28) and the addition rule of probability, an EPPF must satisfy

$$p(1) = 1 \text{ and } p(\boldsymbol{n}) = \sum_{j=1}^{k(\boldsymbol{n})+1} p(\boldsymbol{n}^{j+}) \qquad (\boldsymbol{n} \in \mathbb{N}^*) \tag{29}$$

Conversely [55], every non-negative, symmetric function p defined on $\mathbb{N}^*$ and satisfying (29) is an EPPF, that is to say the EPPF of some exchangeable sequence (X_n).

Proposition 13 *Corresponding to each pair (p, ν), where p is an* EPPF *and ν is a diffuse probability distribution, there is a unique distribution for a species sampling sequence (X_n) such that p is the* EPPF *of (X_n) and ν is the distribution of X_1.*

Proof. For each n the EPPF p of an exchangeable sequence (X_n) determines the probability of each event of the form displayed in (28). But from the prediction rule (15)–(16), given such an event, the common value $\tilde{X}_j$ of X_ℓ for $\ell \in A_j$ has distribution ν, and these $\tilde{X}_j$ are independent for $1 \leq j \leq k$. Thus the joint distribution of $(X_1, \ldots, X_n)$ is determined for every n by (p, ν), which proves the uniqueness claim. Given a pair (p, ν), such a sequence (X_n) is constructed by assigning the values of an i.i.d(ν) sequence $(\tilde{X}_j)$ to the classes of an independent exchangeable random partition of $\mathbb{N}$ defined by p. $\Box$

According to Proposition 11, the random distribution F governing a species sampling sequence (X_n) can be recovered almost surely from (X_n). Thus a pair (p, ν) as above determines the finite-dimensional distributions of a random distribution F such that a sample (X_n) from F has EPPF p and each X_n with distribution ν.

3.3 Ranked frequencies

Kingman's theory of random partitions [46] sets up a one-one correspondence between EPPF's p and distributions for a decreasing sequence of random variables (P_i) with $P_i \geq 0$ and $\sum_i P_i \leq 1$. The random distribution F corresponding to (p, ν) is then constructed via (22). Let $\mathcal{P}$ denote the set of EPPF's $p : \mathbb{N}^* \to [0, 1]$, and give $\mathcal{P}$ the topology of pointwise convergence. The set $\mathcal{P}$ is convex and compact, in fact a *simplex*: as shown by Kingman, the extreme p are those corresponding to a deterministic decreasing sequence of frequencies (P_i).

For a proper sequence (P_i) there is a formula for the corresponding EPPF which follows easily from (28) and (22) by conditioning on (P_i) and $(\hat{X}_i)$:

$$p(n_1, \ldots, n_k) = \sum_{(i_1, \ldots, i_k)} \mathbb{E}\left(\prod_{j=1}^{k} P_{i_j}^{n_j} \right) \tag{30}$$

where the sum is over all sequences of distinct positive integers $(i_1, \ldots, i_k)$, and $\mathbb{E}$ stands for expectation with respect to the underlying probability distribution $\mathbb{P}$. In principle this formula determines the correspondence between the EPPF and the distribution of ranked frequencies (P_i) in a proper species sampling model. But it gives little hint of how to arrange the distribution of frequencies to produce models with a simple EPPF.

3.4 Prediction rules

Consider now the functions p_j defined on $\mathbb{N}^*$ which determine the prediction rule (16) of a species sampling sequence (X_n). From formula (28) and Bayes' rule, these functions are expressed as follows in terms of the EPPF p of (X_n):

$$p_j(\boldsymbol{n}) = \frac{p(\boldsymbol{n}^{j+})}{p(\boldsymbol{n})} \text{ for } 1 \leq j \leq k(\boldsymbol{n}) + 1 \text{ provided } p(\boldsymbol{n}) > 0. \tag{31}$$

It is now clear from Proposition 13 that the statement of Proposition 11 can be sharpened as follows:

Theorem 14 *Given a diffuse probability distribution ν and a sequence of functions $(p_j, j = 1, 2, \ldots)$ defined on $\mathbb{N}^*$ and satisfying* (18)*, let (X_n) be governed by the prediction rule* (15)–(16)*. The sequence (X_n) is exchangeable iff there exists a non-negative, symmetric function p defined on $\mathbb{N}^*$ such that* (31) *holds. Then (X_n) is a sample from F as in Proposition 11, and the* EPPF *of (X_n) is the unique non-negative symmetric function p such* (31) *holds and $p(1) = 1$.*

Example 15 *The Blackwell-MacQueen Urn Scheme.* Fix $\theta > 0$. It is easily checked that the functions p_j displayed in (19) are of the form (31) for the function $p = p_{(\theta)}$ defined on $\mathbb{N}^*$ by

$$p_{(\theta)}(n_1, \ldots, n_k) := \frac{\theta^{k-1} \prod_{i=1}^{k} (n_i - 1)!}{[1+\theta]_{n-1}} \tag{32}$$

where $n = \sum_i n_i$ and $[x]_m = \prod_{j=1}^{m} (x + j - 1)$. Since this function is symmetric, and $p_{(\theta)}(1) = 1$, Theorem 14 implies that (X_n) generated by the Blackwell-MacQueen urn scheme is governed by the species sampling model corresponding to the pair $(p_{(\theta)}, \nu)$. Comparison with Theorem 2 identifies $p_{(\theta)}$ as the EPPF of a sample from DIRICHLET$(\theta\nu)$ for any diffuse distribution ν, a result due to Antoniak [3].

Example 16 *The Two-Parameter Model* [55]. Consider the prediction rule (15)–(16) defined by the sequence of functions

$$p_j(n_1, \ldots, n_k) = \frac{n_j - \alpha}{n + \theta} 1(1 \leq j \leq k) + \frac{\theta + k\alpha}{n + \theta} 1(j = k + 1) \tag{33}$$

where α and θ are two real parameters. To ensure that all relevant probabilities are non-negative and that the rule is not degenerate, it must be supposed that either

$$\alpha = -\kappa < 0 \text{ and } \theta = m\kappa \text{ for some } \kappa > 0 \text{ and } m = 2, 3 \ldots \tag{34}$$

or

$$0 \leq \alpha < 1 \text{ and } \theta > -\alpha \tag{35}$$

This prediction rule satisfies (31) for the function $p = p_{(\alpha,\theta)}$ defined by

$$p_{(\alpha,\theta)}(n_1, \ldots, n_k) = \frac{\left(\prod_{\ell=1}^{k-1} (\theta + \ell\alpha)\right) \left(\prod_{i=1}^{k} [1-\alpha]_{n_i - 1}\right)}{[1+\theta]_{n-1}} \tag{36}$$

where $n = \sum_i n_i$ and $[x]_m = \prod_{j=1}^{m}(x + j - 1)$. Since $p_{(\alpha,\theta)}(1) = 1$ and $p_{(\alpha,\theta)}$ is symmetric, Theorem 14 shows that (X_n) defined by this prediction rule is exchangeable, hence a species sampling sequence. The case (34) corresponds to sampling from $F = \sum_{i=1}^{m} P_i \hat{X}_i$ where $(P_1, \ldots, P_m)$ has a symmetric Dirichlet distribution with m parameters equal to κ, and the $\hat{X}_i$ are i.i.d. with distribution ν. This is just Fisher's model, described in Section 2, with m species identified by i.i.d.(ν) tags. In this model, the number of species K_n in a sample of size n remains bounded above by m and is eventually equal to m as $n \to \infty$. Passing to the limit as $m \to \infty$ and $\kappa \to 0$ for fixed $\theta = m\kappa$, the prediction rule and the EPPF for Fisher's model converge to the Blackwell-MacQueen rule and the EPPF for sampling from a Dirichlet$(\theta\nu)$ prior, which is the special case of the two-parameter model with $\alpha = 0$ and $\theta > 0$. In this model, K_n is a sum of independent indicator variables, which implies $K_n \sim \theta \log n$ almost surely and K_n is asymptotically normal [48]. In the model with $0 < \alpha < 1$ and $\theta > -\alpha$ the sequence (K_n) is an inhomogeneous Markov chain such that $K_n \sim S n^{\alpha}$ almost surely, for a random variable S with a continuous density on $(0, \infty)$ depending on (α, θ). See [56, 60, 54] for this and other asymptotic results for the two-parameter model with $0 < \alpha < 1$, which follow from an extension to this case of the Poisson representation of Theorem 3.

3.5 The sampling formula

For $n = 1, 2, \ldots$, define a vector $\boldsymbol{C}_n = (C_{1n}, \ldots, C_{nn})$ of non-negative integer counts by

$$C_{in} = \sum_{j=1}^{n} 1(N_{jn} = i) \tag{37}$$

So C_{in} represents the number of species that appear exactly i times among $X_1, \ldots, X_n$. By definition $\sum_i iC_{in} = n$, and $\sum_i C_{in} = K_n$. The vector $\boldsymbol{C}_n$ is a standard coding of the *partition of* n induced by $X_1, \ldots, X_n$. Instead of working with the EPPF p as above, Kingman [46] worked with the function

$$p^{\star}(\boldsymbol{m}) = \mathbb{P}(\boldsymbol{C}_n = \boldsymbol{m}) \tag{38}$$

defined for finite vectors of non-negative integers $\boldsymbol{m} = (m_1, \ldots, m_n)$. For fixed n, as $\boldsymbol{m} = (m_1, \ldots, m_n)$ ranges over all such vectors of length n with $\Sigma\, im_i = n$, this function $p^{\star}(\boldsymbol{m})$ defines a probability distribution over partitions of n. In terms of species sampling, $p^{\star}(\boldsymbol{m})$ is the probability that in a sample of size n there are m_1 species with a single representative in the sample, and m_2 species with two representatives in the sample, and so on. By an elementary counting argument, the number of partitions of the set $[n]$

that contain m_1 singleton sets, m_2 doubletons, and so on, is

$$\#(\boldsymbol{m}) := \frac{n!}{\prod_{i=1}^{n}(i!)^{m_i} m_i!} \tag{39}$$

It follows that

$$p^{\star}(\boldsymbol{m}) = \#(\boldsymbol{m}) p^{\circ}(\boldsymbol{m}) \tag{40}$$

where

$$p^{\circ}(m_1, \ldots, m_n) = p(n_1, \ldots, n_k) \tag{41}$$

for every sequence $(n_1, \ldots, n_k)$ such that

$$m_i = \sum_{\ell=1}^{k} 1(n_\ell = i) \text{ for all } 1 \le i \le k,$$

due to the symmetry of p. To illustrate, for (X_n) governed by the two-parameter model with EPPF $p = p_{(\alpha,\theta)}$ as in (36), it follows that $p^{\star}(\boldsymbol{m}) = p^{\star}_{(\alpha,\theta)}(\boldsymbol{m})$ is given by the following formula [55]: for $\boldsymbol{m} = (m_1, \ldots, m_n)$ with $\Sigma\, m_i = k$, and $\Sigma\, i m_i = n$,

$$p^{\star}_{(\alpha,\theta)}(\boldsymbol{m}) = n! \, \frac{\left(\prod_{\ell=1}^{k-1}(\theta + \ell\alpha)\right)}{[\theta+1]_{n-1}} \prod_{i=1}^{n} \left(\frac{[1-\alpha]_{i-1}}{i!}\right)^{m_i} \frac{1}{m_i!} \tag{42}$$

For $\alpha = 0, \theta > 0$ this is the *Ewens sampling formula* [23] which has found extensive applications in population genetics [47, 24, 25]. Antoniak [3] showed that the distribution of the partition of n generated by sampling from a DIRICHLET$(\theta\nu)$ prior for a diffuse ν is governed by this formula. The case $\alpha = -\kappa < 0$ and $\theta = m\kappa$ gives the distribution of the partition of n generated by Fisher's m-species model, corresponding to sampling from a symmetric Dirichlet prior on m points. The formula in this case was found by Watterson [70].

3.6 The frequencies in order of appearance

Consider now the infinite sequence $\tilde{\boldsymbol{P}} = (\tilde{P}_1, \tilde{P}_2, \ldots)$ of frequencies in order of appearance obtained from a species sampling sequence (X_n), as in (19). According to Theorem 6 of [55], for each $n = 1, 2, \ldots$ the conditional distribution of X_{n+1} given $X_1, \ldots, X_n$ and $\tilde{\boldsymbol{P}}$ is given by

$$P(X_{n+1} \in \cdot \,|\, X_1, \ldots, X_n, \tilde{\boldsymbol{P}}) = \sum_{j=1}^{K_n} \tilde{P}_j 1(\tilde{X}_j \in \cdot) + (1 - \sum_{j=1}^{K_n} \tilde{P}_j\,)\nu(\cdot) \tag{43}$$

For $\boldsymbol{n} = (n_1, \ldots, n_k) \in \mathbb{N}^*$, this formula shows that the conditional probability of the event displayed in (28) given $\tilde{\boldsymbol{P}}$ is $\Pi(\boldsymbol{n}, \tilde{\boldsymbol{P}})$ where

$$\Pi(\boldsymbol{n}, \tilde{\boldsymbol{P}}) = \left(\prod_{i=1}^{k-1} \left[\tilde{P}_i^{n_i - 1} (1 - \sum_{j=1}^{i} \tilde{P}_j) \right] \right) \tilde{P}_k^{n_k - 1} \tag{44}$$

It follows that the EPPF p of (X_n) and the distribution of the sequence $\tilde{\boldsymbol{P}}$ determine each other via the formula

$$p(\boldsymbol{n}) = \mathbb{E}[\Pi(\boldsymbol{n}, \tilde{\boldsymbol{P}})] \tag{45}$$

See [55] for details. In view of McCloskey's Theorem 5, it is natural to represent the frequencies $\tilde{P}_j$ in the form

$$\tilde{P}_j = \left[\prod_{i=1}^{j-1} (1 - \tilde{W}_i) \right] \tilde{W}_j \tag{46}$$

for a sequence of random variables $(\tilde{W}_j)$ with $0 \leq \tilde{W}_j \leq 1$. The formula (45) then becomes

$$\Pi(\boldsymbol{n}, \tilde{\boldsymbol{P}}) = \prod_{i=1}^{k} \tilde{W}_i^{n_i - 1} (1 - \tilde{W}_i)^{n_{i+1} + \cdots + n_k} \tag{47}$$

Keep in mind that the $\tilde{W}_i$ are subject not only to $0 \leq \tilde{W}_i \leq 1$, but also to the more subtle *symmetry constraint* that $\mathbb{E}[\Pi(\boldsymbol{n}, \tilde{\boldsymbol{P}})]$ is a symmetric function of $\boldsymbol{n}$. It is natural to look first at models in which the $\tilde{W}_j$ are independent. But the choice is severely limited by the symmetry constraint. Assuming for simplicity that $0 < \tilde{W}_j < 1$ there is the following generalization of McCloskey's Theorem 5:

Theorem 17 [58, 55] *A species sampling sequence (X_n) is such that the frequencies in order of appearance $(\tilde{P}_j)$ are of the form* (46) *for independent random variables $(\tilde{W}_j)$ with $0 < \tilde{W}_j < 1$ a.s. iff (X_n) is governed by the two-parameter prediction rule* (33) *for some $0 \leq \alpha < 1$ and $\theta > -\alpha$. Then $\tilde{W}_j$ has beta$(1 - \alpha, \theta + j\alpha)$ distribution for all j.*

3.7 The updating rule

The argument leading to (45), combined with Bayes' rule, yields the following updating rule for species sampling:

Theorem 18 *Suppose (X_n) is a species sampling sequence. For each $n = 1, 2, \ldots$ the conditional distribution of F given $(X_1, \ldots, X_n)$ is determined by* *the following conditional distribution given $(X_1, \ldots, X_n)$ of the two sequences $(\tilde{X}_j)$ and $\tilde{\boldsymbol{P}} := (\tilde{P}_j)$ which determine F via (20):*

(i) *the $\tilde{X}_j$ for $1 \le j \le K_n$ are measurable functions of $(X_1, \ldots, X_n)$*

(ii) *the $\tilde{X}_j$ for $j > K_n$ are i.i.d (ν)*

(iii) *independent of the all the $\tilde{X}_j$, the sequence $\tilde{\boldsymbol{P}}$ has conditional distribution specified by the following formula: for all non-negative product measurable functions f*

$$\mathbb{E}[f(\tilde{\boldsymbol{P}})|X_1, \ldots, X_n, \boldsymbol{N}_n = \boldsymbol{n}] = \frac{\mathbb{E}[f(\tilde{\boldsymbol{P}})\Pi(\boldsymbol{n}, \tilde{\boldsymbol{P}})]}{p(\boldsymbol{n})} \tag{48}$$

where $\Pi(\boldsymbol{n}, \tilde{\boldsymbol{P}})$ is defined by (43), and $p(\boldsymbol{n}) = \mathbb{E}[\Pi(\boldsymbol{n}, \tilde{\boldsymbol{P}})]$ is the EPPF *of (X_n).*

Note the following special cases of formula (48), which relate this formula to the prediction probability functions p_j via (31), and can be read directly from (16) and (43):

$$\mathbb{E}[\tilde{P}_j|X_1, \ldots, X_n, \boldsymbol{N}_n = \boldsymbol{n}] = p_j(\boldsymbol{n}) \qquad (1 \le j \le k(\boldsymbol{n})) \tag{49}$$

and for $\tilde{R}_k := 1 - \sum_{i=1}^k \tilde{P}_i$,

$$\mathbb{E}[\tilde{R}_{K_n}|X_1, \ldots, X_n, \boldsymbol{N}_n = \boldsymbol{n}] = p_{k(\boldsymbol{n})+1}(\boldsymbol{n}) \tag{50}$$

Here the random variable $\tilde{R}_{K_n}$ represents the proportion in the total population of all species unobserved in the sample $X_1, \ldots, X_n$.

To illustrate, consider sampling from F corresponding to the two-parameter prediction rule in Example 16.

Definition 19 Say a random discrete distribution F has (α, θ, ν)-*distribution* if a sample (X_n) from F is governed by the model of Example 16 determined by real parameters α and θ subject to either (34) or (35), and a diffuse measure ν.

That is to say, according to Theorem 17, F has the same distribution as $\sum_j \tilde{P}_j \delta(\tilde{X}_j)$ where the $\tilde{P}_j$ are given by (46) for $\tilde{W}_i$ that are independent with beta$(1-\alpha, \theta + i\alpha)$ distributions, independent also of the i.i.d(ν) sequence $(\tilde{X}_j)$. In particular, the $(0, \theta, \nu)$ distribution is DIRICHLET$(\theta\nu)$. To be definite, it will be assumed that $0 \le \alpha < 1$ and $\theta > -\alpha$. But the following results hold just as well for the range of parameters (34) corresponding to Fisher's model, provided attention is restricted to $\tilde{W}_i$ for $1 \le i \le m$ and it is understood that $\tilde{W}_m = 1$.

If (X_n) is a sample from F with the (α, θ, ν)-distribution, then the conditional distribution of $\tilde{\boldsymbol{P}}$ given $(X_1, \ldots, X_n)$ can be made more explicit as follows. Using the expression (47) for $\Pi(\boldsymbol{n}, \tilde{\boldsymbol{P}})$, it follows from (48) that given $(X_1, \ldots, X_n)$ with $\boldsymbol{N}_n = (n_1, \ldots, n_k)$, the $\tilde{W}_i$ are independent, with

$$\text{beta}(n_i - \alpha, \theta + i\alpha + \textstyle\sum_{j=i+1}^{k} n_j) \text{ distribution for } 1 \leq i \leq k, \text{ and} \tag{51}$$

$$\text{beta}(1 - \alpha, \theta + i\alpha) \text{ distribution for } i > k. \tag{52}$$

This amounts to the following updating rule:

Corollary 20 *If (X_n) is a sample from a random distribution F with the (α, θ, ν)-distribution, then conditionally given $X_1, \ldots, X_n$ with k distinct values $\tilde{X}_j$ for $1 \leq j \leq k$, and n_j values X_i equal to $\tilde{X}_j$ for each $1 \leq j \leq k$,*

$$F = \sum_{j=1}^{k} \tilde{P}_j \delta(\tilde{X}_j) + \tilde{R}_k F_k \tag{53}$$

where $(\tilde{P}_1, \ldots, \tilde{P}_k, \tilde{R}_k)$ has DIRICHLET$(n_1 - \alpha, \ldots, n_k - \alpha, \theta + k\alpha)$ *distribution, independently of the random distribution F_k, which has $(\alpha, \theta + k\alpha, \nu)$-distribution.*

As checks, take expectations and use the formulae (49) and (50) to recover the (α, θ) prediction rule (33). Also, for $\alpha = 0$ it is easily verified that this updating rule reduces to the Dirichlet updating rule of Theorem 1.

4 Random Permutations

It is easily seen that there exists a unique probability distribution for a sequence of random permutations $(\sigma_n, n = 1, 2, \cdots)$ such that

(i) σ_n is a uniformly distributed random permutation of $[n]$ for each n;

(ii) for each n, if σ_n is written as a product of cycles, then σ_{n-1} is derived from σ_n by deletion of element n from its cycle.

For example, using standard cycle notation for permutations,

if $\sigma_5 = (134)(25)$ then $\sigma_4 = (134)(2)$;
if $\sigma_5 = (134)(2)(5)$ then $\sigma_4 = (134)(2)$.

The combinatorial basis of the above observation appears in Greenwood [32]. Lester Dubins and I devised the following *Chinese restaurant construction* of such a sequence (σ_n), which is mentioned in [1, (11.19)]. Suppose people numbered $1, 2, \cdots$ arrive in an initially empty restaurant with an unlimited number of circular tables $T_1, T_2, \cdots$, each capable of seating an unlimited number of people. Person 1 sits at table T_1. For $n \geq 1$ suppose inductively that n people have already entered the restaurant, and are seated in some arrangement, with at least one person at each of the tables T_j for $1 \leq j \leq k$ say, where k is the number of tables occupied by the first n people to arrive. Let person $n+1$ choose with equal probability to sit at any of the following $n+1$ places: to the left of person j for some $1 \leq j \leq n$, or alone at table T_{k+1}. Define $\sigma_n : [n] \to [n]$ by $\sigma_n(i) = j$ if person j is seated immediately to the left of person i after n people have entered. The sequence (σ_n) then has features (i) and (ii) above by a simple induction.

Suppose now that independently of the sequence (σ_n) the table T_j is painted a random color $\tilde{X}_j$, where the $\tilde{X}_j$ are i.i.d. random variables with some arbitrary probability distribution ν over a spectrum S of possible colors. Let X_n denote the color of the table occupied by the nth person to arrive. By construction, the sequential development of (X_n) is exactly that described by the Blackwell-MacQueen urn scheme for $\mu = \nu$. Consequently,

$$(X_n) \text{ is a sample from } F \text{ with } \textsc{dirichlet}(\nu) \text{ distribution} \tag{54}$$

where F is the limiting empirical distribution of $X_1, \cdots, X_n$ as $n \to \infty$. Assuming ν is diffuse, different tables have different colors almost surely. Then by construction, the following three partitions of $[n]$ are almost surely identical: the partition of $[n]$ induced by $X_1, \cdots, X_n$, the partition of $[n]$ defined by the way the first n customers are distributed among tables, and the partition of $[n]$ induced by the cycles of the uniformly distributed random permutation σ_n.

To construct F with $\textsc{dirichlet}(\theta\nu)$ distribution for arbitrary $\theta > 0$, let people enter the restaurant exactly as before, but suppose that given the seating arrangement of the first n people at tables $T_1, \cdots, T_k$ say, person $n+1$ chooses to sit the left of person j with equal probability $1/(n+\theta)$ for each $1 \leq j \leq n$, and to sit alone at table T_{k+1} with probability $\theta/(n+\theta)$. Now the sequence of colors (X_n) is governed by the Blackwell-MacQueen urn scheme for $\mu = \theta\nu$, so

$$(X_n) \text{ is a sample from } F \text{ with } \textsc{dirichlet}(\theta\nu) \text{ distribution} \tag{55}$$

In this construction the number K_n of cycles of σ_n is represented as

$$K_n = Y_1 + \cdots + Y_n \tag{56}$$

where Y_m is the indicator of the event that customer m chooses to sit alone at a new table, and by construction the Y_m are independent Bernoulli $(\theta/(\theta + m - 1))$ variables. In terms of (X_n) constructed as a sample from Dirichlet$(\theta\nu)$ for a diffuse ν, Y_m is the indicator of the event that X_m does not equal X_j for any $1 \leq j \leq m-1$, and K_n is the number of distinct values observed among $X_1, \cdots, X_n$. The simple structure of the representation (56) in this setting can be read immediately from the Blackwell-MacQueen urn scheme. As noted by Korwar-Hollander [48], by application of standard limit theorems for sums of independent random variables, this leads to a law of large numbers and a central limit theorem governing the $\mathbb{P}_\theta$ asymptotic behaviour of K_n for large n. In particular, for $\theta = 1$, we recover the result of Goncharov [31] regarding the asymptotic normality for large n of the distribution of the number of cycles of σ_n, a uniformly distributed random permutation of $[n]$. A number of other asymptotic results for random permutations, originally obtained by other methods [65, 68], can also derived from the Chinese restaurant construction. See also [5, 4, 6, 13, 35, 39, 57] for related developments.

The Chinese restaurant construction can also be extended to the more general species sampling setting of Section 3. See Kerov [44, §4.2] for a remarkable connection between these ideas and the theory of interlacing measures and the Markov transform.

References

[1] D.J. Aldous. Exchangeability and related topics. In P.L. Hennequin, editor, *École d'Été de Probabilités de Saint-Flour XII, Springer Lecture Notes in Mathematics, Vol. 1117*. Springer-Verlag, 1985.

[2] F. J. Anscombe. Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika*, 37:358–382, 1950.

[3] C. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.*, 2:1152–1174, 1974.

[4] R. Arratia, A.D. Barbour, and S. Tavaré. Poisson process approximations for the Ewens sampling formula. *Ann. Appl. Prob.*, 2:519–535, 1992.

[5] R. Arratia and S. Tavaré. The cycle structure of random permutations. *Ann. Prob.*, 20:1567–1591, 1992.

[6] A. D. Barbour. Refined approximations for the Ewens sampling formula. *Rand. Struct. Alg.*, 3:267–276, 1992.

[7] R.H. Berk and I.R. Savage. Dirichlet processes produce discrete measures: an elementary proof. In *Contributions to Statistics. Jaroslav Hajek Memorial Volume*, pages 25–31. Academia, North Holland, Prague, 1979.

[8] P. Billingsley. On the distribution of large prime factors. *Period. Math. Hungar.*, 2:283–289, 1972.

[9] D. Blackwell. Discreteness of Ferguson selections. *Ann. Statist.*, 1:356–358, 1973.

[10] D. Blackwell and J.B. MacQueen. Ferguson distributions via Pólya urn schemes. *Ann. Statist.*, 1:353–355, 1973.

[11] J. Bunge and M. Fitzpatrick. Estimating the number of species: A review. *Journal of the American Statistical Association*, 88:364 – 373, 1993.

[12] R.J. Connor and J.E. Mosimann. Concepts of independence for proportions with a generalization of the Dirichlet distribution. *J. Am. Statist. Assoc.*, 64:194–206, 1969.

[13] P. Diaconis and J. Kemperman. Some new tools for Dirichlet priors. In J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, editors, *Bayesian Statistics*, pages 95–104. Oxford Univ. Press, 1995.

[14] K. Dickman. On the frequency of numbers containing prime factors of a certain relative magnitude. *Ark. Mat. Astronomi och Fysik*, 22:1–14, 1930.

[15] K. Doksum. Tailfree and neutral random probabilities and their posterior distributions. *The Annals of Probability*, 2:183 – 201, 1974.

[16] P. Donnelly and G. Grimmett. On the asymptotic distribution of large prime factors. *J. London Math. Soc. (2)*, 47:395–404, 1993.

[17] P. Donnelly and P. Joyce. Continuity and weak convergence of ranked and size-biased permutations on the infinite simplex. *Stochastic Processes and their Applications*, 31:89 – 103, 1989.

[18] S. Engen. *Stochastic Abundance Models with Emphasis on Biological Communities and Species Diversity.* Chapman and Hall Ltd., 1978.

[19] S. N. Ethier and T.G. Kurtz. The infinitely-many-neutral-alleles diffusion model. *Advances in Applied Probability*, 13:429 – 452, 1981.

[20] S.N. Ethier. The distribution of the frequencies of age-ordered alleles in a diffusion model. *Adv. Appl. Prob.*, 22:519–532, 1990.

[21] S.N. Ethier and T.G. Kurtz. Fleming-Viot processes in population genetics. *SIAM Journal on Control and Optimization*, 31:345–386, 1993.

[22] S.N. Ethier and T.G. Kurtz. Convergence to Fleming–Viot processes in the weak atomic topology. *Stoch. Proc. Appl.*, 54:1–27, 1994.

[23] W.J. Ewens. The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.*, 3:87 – 112, 1972.

[24] W.J. Ewens. Population genetics theory – the past and the future. In S. Lessard, editor, *Mathematical and Statistical Problems in Evolution.* University of Montreal Press, Montreal, 1988.

[25] W.J. Ewens and S. Tavaré. The Ewens sampling formula. To appear in *Multivariate Discrete Distributions* edited by N.S. Johnson, S. Kotz, and N. Balakrishnan, 1995.

[26] J. Fabius. Asymptotic behavior of Bayes estimates. *Ann. Math. Statist.*, 35:846–856, 1964.

[27] T.S. Ferguson. A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1:209–230, 1973.

[28] T.S. Ferguson. Prior distributions on spaces of probability measures. *Ann. Statist.*, 2:615–629, 1974.

[29] R.A. Fisher, A.S. Corbet, and C.B. Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Animal Ecol.*, 12:42–58, 1943.

[30] D. Freedman. On the asymptotic behavior of Bayes estimates in the discrete case. *Ann. Math. Statist.*, 34:1386–1403, 1963.

[31] V. Gončharov. On the field of combinatory analysis. *Amer. Math. Soc. Transl.*, 19:1–46, 1962.

[32] R.E. Greenwood. The number of cycles associated with the elements of a permutation group. *Amer. Math. Monthly*, 60:407–409, 1953.

[33] P. Halmos. Random alms. *Ann. Math. Stat.*, 15:182–189, 1944.

[34] J.C. Hansen. Order statistics for decomposable combinatorial structures. *Rand. Struct. Alg.*, 5:517–533, 1994.

[35] F. M. Hoppe. Pólya-like urns and the Ewens sampling formula. *Journal of Mathematical Biology*, 20:91 – 94, 1984.

[36] F. M. Hoppe. Size-biased filtering of Poisson-Dirichlet samples with an application to partition structures in genetics. *Journal of Applied Probability*, 23:1008 – 1012, 1986.

[37] F. M. Hoppe. The sampling theory of neutral alleles and an urn model in population genetics. *Journal of Mathematical Biology*, 25:123 – 159, 1987.

[38] T. Ignatov. On a constant arising in the theory of symmetric groups and on Poisson-Dirichlet measures. *Theory Probab. Appl.*, 27:136–147, 1982.

[39] P. Joyce and S. Tavaré. Cycles, permutations and the structure of the yule process with immigration. *Stochastic Process. Appl.*, 25:309–314, 1987.

[40] O. Kallenberg. A canonical representation of symmetrically distributed random measures. In P. Jagers and L. Rade, editors, *Mathematics and Statistics: Essays in Honour of Harald Bergström*, pages 41–48. Teknologtryck, Göteborg, Sweden, 1973.

[41] O. Kallenberg. Canonical representations and convergence criteria for processes with interchageable increments. *Z. Wahrsch. Verw. Gebiete*, 27:23–36, 1973.

[42] O. Kallenberg. On symmetrically distributed random measures. *Trans. Amer. Math. Soc.*, 202:105–121, 1975.

[43] S. Kerov. Coherent random allocations and the Ewens-Pitman formula. PDMI Preprint, Steklov Math. Institute, St. Petersburg, 1995.

[44] S. Kerov. Interlacing measures. Technical Report 1116-96, Laboratoire Bordelais de Recherche en Informatique, 1996.

[45] J. F. C. Kingman. Random discrete distributions. *J. Roy. Statist. Soc. B*, 37:1–22, 1975.

[46] J. F. C. Kingman. The representation of partition structures. *J. London Math. Soc.*, 18:374–380, 1978.

[47] J.F. C. Kingman. Exchangeability and the evolution of large populations. In G. Koch and F. Spizzichino, editors, *Exchangeability in Probability and Statistics.* North-Holland Publishing Company, 1982.

[48] R. M. Korwar and M. Hollander. Contributions to the theory of Dirichlet processes. *Ann. Prob.*, 1:705–711, 1973.

[49] R. D. Mauldin, W. D. Sudderth, and S. C. Williams. Polya trees and random distributions. *Annals of Statistics*, 20:1203–1221, 1992.

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

[51] G. P. Patil and C. Taillie. Diversity as a concept and its implications for random communities. *Bull. Int. Stat. Inst.*, XLVII:497 – 515, 1977.

[52] M. Perman. Order statistics for jumps of normalized subordinators. *Stoch. Proc. Appl.*, 46:267–281, 1993.

[53] M. Perman, J. Pitman, and M. Yor. Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields*, 92:21–39, 1992.

[54] J. Pitman. Partition structures derived from Brownian motion and stable subordinators. Technical Report 346, Dept. Statistics, U.C. Berkeley, 1992. To appear in *Bernoulli.*

[55] J. Pitman. Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102:145–158, 1995.

[56] J. Pitman. Species sampling models. In preparation, 1995.

[57] J. Pitman. Probabilistic bounds on the coefficients of polynomials with only real zeros. Technical Report 453, Dept. Statistics, U.C. Berkeley, 1996. To appear in *J. Comb. Theory A.*

[58] J. Pitman. Random discrete distributions invariant under size-biased permutation. To appear in *Advances in Applied Probability*, 1996.

[59] J. Pitman and M. Yor. Arcsine laws and interval partitions derived from a stable subordinator. *Proc. London Math. Soc. (3)*, 65:326–356, 1992.

[60] J. Pitman and M. Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. Technical Report 433, Dept. Statistics, U.C. Berkeley, 1995. To appear in *The Annals of Probability.*

[61] J.-M. Rolin. Some useful properties of the Dirichlet process. Institut de Statistique, Université Catholique de Louvain, Discussion Paper No. 9202, 1992.

[62] J.-M. Rolin. On the distribution of jumps of the Dirichlet Process. Institut de Statistique, Université Catholique de Louvain, Discussion Paper No. 9302, 1993.

[63] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[64] J. Sethuraman and R. C. Tiwari. Convergence of Dirichlet measures and the interpretation of their parameter. *Statistical Decision Theory and Related Topics III*, 2:305–315, 1982.

[65] L.A. Shepp and S.P. Lloyd. Ordered cycle lengths in a random permutation. *Trans. Amer. Math. Soc.*, 121:340–357, 1966.

[66] R.C. Tiwari and R.C. Tripathi. Nonparametric Bayes estimation of the probability of discovering a new species. *Communications in Statistics, Part A–Theory and Methods*, 18:305, 1989.

[67] A. M. Vershik. The asymptotic distribution of factorizations of natural numbers into prime divisors. *Soviet Math. Dokl.*, 34:57–61, 1986.

[68] A.M. Vershik and A.A. Shmidt. Limit measures arising in the theory of groups, I. *Theor. Prob. Appl.*, 22:79–85, 1977.

[69] A.M. Vershik and A.A. Shmidt. Limit measures arising in the theory of symmetric groups, II. *Theor. Prob. Appl.*, 23:36–49, 1978.

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

[71] S.L. Zabell. The continuum of inductive methods revisited. In J. Earman and J. Norton, editors, *The Cosmos of Science*, Pittsburgh-Konstanz Series in the Philosophy and History of Science. University of Pittsburgh Press/Universitatsverlag Konstanz, 1996. To appear.

Department of Statistics
University of California
367 Evans Hall # 3860
Berkeley, CA 94720-3860