Ann. Inst. Statist. Math.
Vol. 57, No. 3, 597–616 (2005)
©2005 The Institute of Statistical Mathematics

UNORDERED AND ORDERED SAMPLE FROM DIRICHLET DISTRIBUTION

THIERRY HUILLET

Laboratoire de Physique Théorique et Modélisation, CNRS-UMR 8089 et Université de Cergy-Pontoise, 5 mail Gay-Lussac, 95031, Neuville sur Oise, France,
e-mail: Thierry.Huillet@ptm.u-cergy.fr

(Received March 2, 2004; revised September 1, 2004)

Abstract. Consider the random Dirichlet partition of the interval into n fragments with parameter $\theta > 0$. Explicit results on the statistical structure of its size-biased permutation are recalled, leading to (unordered) Ewens and (ordered) Donnelly-Tavaré-Griffiths sampling formulae from finite Dirichlet partitions. We use these preliminary statistical results on frequencies distribution to address the following sampling problem: what are the intervals between new sampled categories when sampling is from Dirichlet populations? The results obtained are in accordance with the ones found in sampling theory from random proportions with GEM(γ) distribution. These can be obtained from Dirichlet model when considering the Kingman limit $n \uparrow \infty$, $\theta \downarrow 0$ while $n\theta = \gamma > 0$.

Key words and phrases: Random discrete distribution, Dirichlet partition, size-biased permutation, GEM, Ewens and Donnelly-Tavaré-Griffiths sampling formulae, intervals between new sampled species.

1. Introduction

The joint distribution of unordered (or ordered) frequencies of a sample from random proportions with Griffiths-Engen-McCloskey, or $GEM(\gamma)$, distribution is the Donnelly-Tavaré-Griffiths formula (or the Ewens sampling formulae). The $GEM(\gamma)$ distribution will be defined later (see also Kingman (1993), Chapter 9, for example).

We first reconsider the same sampling problems and formulae when sampling is from random proportions with Dirichlet $D_n(\theta)$ distribution, hence with a finite number n of fragments in the partition. The usual Ewens and Donnelly-Tavaré-Griffiths sampling formulae can be found when passing to the Kingman limit $n \uparrow \infty$, $\theta \downarrow 0$, $n\theta = \gamma > 0$. These preliminary results which are recalled are used to study the intervals between consecutive categories and number of distinct categories when sampling is from $D_n(\theta)$. The results obtained are in accordance with particular cases of those of Yamato-Sibuya-Nomachi when sampling is from $GEM(\gamma)$, while passing to the Kingman limit.

The organization of this manuscript is the following. Consider the random Dirichlet partition of the interval into n fragments with parameter $\theta > 0$. Elementary properties of its $D_n(\theta)$ distribution are first recalled in Section 2. Explicit results on the law of its size-biased permutation are also recalled there. These are useful when considering the Donnelly-Tavaré-Griffiths sampling formula from Dirichlet partitions in subsequent Subsection 3.3, as given in Theorem 3.3. A size-biased permutation of the fragments

sizes is the one obtained in a size-biased sampling process without replacement from a Dirichlet partition. The main points which we recall are the following: in Lemma 2.1, its residual allocation model structure is recalled. In Lemma 2.2, the order in which the consecutive fragments are visited is considered. In Theorem 2.1, we recall the joint law of the size-biased permutation fragments sizes explicitly. Using this, it is recalled in Corollary 2.1 that consecutive fragments in the size-biased permuted partition are arranged in stochastic descending order.

Section 3 recalls the (unordered) Ewens and (ordered) Donnelly-Tavaré-Griffiths sampling formulae when sampling is from finite Dirichlet partitions. In more details, Subsection 3.1 is devoted to the first Ewens sampling formula when sampling is from Dirichlet partition $D_n(\theta)$. Here the order in which sequentially sampled species arise is irrelevant. Subsection 3.2 concerns the second Ewens sampling formula under the same hypothesis (as a problem of random partitioning of the integers) and Subsection 3.3 deals with the finite Dirichlet version of the Donnelly-Tavaré-Griffiths sampling formula. Here, the order of appearance of sampled species is taken into account. Main results are displayed in Theorems 3.1, 3.2 and 3.3 for each of the problems alluded to. As corollaries to these theorems, the usual well-known sampling formulae can be deduced in each case when sampling is from GEM distribution which is the limiting version of the size-biased permutation of Dirichlet partitions in the sense of Kingman.

The main body of our new sampling results is in Section 4. We use the previous statistical results on frequencies distribution to address the following sampling problem: what are the intervals between new sampled categories until exhaustion of the list, when sampling is from Dirichlet populations? One intuitively expects these intervals to be increasing while approaching complete exhaustion. Our main results on the statistical structure of these intervals are summarized in Theorems 4.1, 4.2 and 4.3, considering both cases of a fixed and then unlimited sample size.

The results obtained are in accordance with the Yamato-Sibuya-Nomachi ones when sampling is from random proportions with $GEM(\gamma)$ distribution. These can naturally be obtained from Dirichlet model when considering the Kingman limit. Results in this direction are displayed in Corollaries 4.1, 4.2 and 4.3 to Theorems 4.1, 4.2 and 4.3, respectively.

2. Size-biased permutation of the Dirichlet distribution

2.1 *Dirichlet partition of the interval*

Consider the following random partition into n fragments of the unit interval. Let $\theta > 0$ be some parameter and assume that the random fragments' sizes $\boldsymbol{S}_n := (S_1, \ldots, S_n)$ (with $\sum_{m=1}^n S_m = 1$) is distributed according to the (exchangeable) Dirichlet $D_n(\theta)$ density function on the simplex, that is to say

$$f_{S_1,\ldots,S_n}(s_1, \ldots, s_n) = \frac{\Gamma(n\theta)}{\Gamma(\theta)^n} \prod_{m=1}^n s_m^{\theta-1} \cdot \delta_{(\sum_{m=1}^n s_m - 1)}. \tag{2.1}$$

Alternatively, the law of $\boldsymbol{S}_n := (S_1, \ldots, S_n)$ is characterized by its joint moment function

$$\boldsymbol{E}\left[\prod_{m=1}^n S_m^{q_m}\right] = \frac{\Gamma(n\theta)}{\Gamma(n\theta + \sum_{m=1}^n q_m)} \prod_{m=1}^n \frac{\Gamma(\theta + q_m)}{\Gamma(\theta)}. \tag{2.2}$$

We shall put $\boldsymbol{S}_n \overset{d}{\sim} D_n(\theta)$ if $\boldsymbol{S}_n$ is Dirichlet distributed with parameter θ.

If this is so, $S_m \stackrel{d}{=} S_n$, $m = 1, \ldots, n$, independently of m and the individual fragments sizes are all identically distributed. Their common density on the interval $(0, 1)$ is a beta$(\theta, (n-1)\theta)$ density.

When $\theta = 1$, the partition model equations (2.1), (2.2) corresponds to the standard uniform partition model of the interval.

In the random division of the interval as in equation (2.1), although all fragments are identically distributed with sizes of order n^{-1}, the smallest fragment's size grows like $n^{-(\theta+1)/\theta}$ while the one of the largest is of order $\frac{1}{n\theta}\log(n\log^{\theta-1} n)$. The smaller θ is, the larger (smaller) the largest (smallest) fragments' size is: hence, the smaller θ is, the more the values of the S_ms are, with high probability, disparate. Smaller the parameter, the size of the largest fragment $S_{(1)}$ tends to dominate the other ones. On the contrary, large values of θ correspond to situations in which the range of fragments' sizes is lower: the fragments' sizes look more homogeneous and distribution equation (2.1) concentrates on its centre. For large θ, the diversity of the partition is small.

Although $\boldsymbol{S}_n$ has a degenerate weak limit when $n \uparrow \infty$, $\theta \downarrow 0$ while $n\theta = \gamma > 0$, this situation is worth being considered (as first noted by Kingman (1975)). Indeed, many interesting statistical features emerge.

2.2 *Size-biased permutation of Dirichlet partitions*

The results on size-biased permutation of Dirichlet distributions presented in this section are not new. When $\theta = 1$, they can be found in Huillet (2003); they were generalized to all $\theta > 0$ in Barrera *et al.* (2005), to solve a problem consisting in computing the search-cost distribution arising from heaps processes. Part of them are reproduced here for the sake of completeness and to make the presentation self-contained. They will prove useful in the sequel.

Assume some observer is sampling the unit interval as follows: drop points at random onto this randomly broken interval and record the corresponding numbers of visited fragments. We shall consider the problem of determining the order in which the various fragments are discovered in such a sampling process. To avoid revisiting the same fragment many times, once it has been discovered, we need to remove it from the population as soon as it has been met in the sampling process. But to do that, the law of its size is needed. Once this is done, after renormalizing the remaining fragments' sizes, we are left with a population of $n-1$ fragments, the sampling of which will necessarily supply a so far undiscovered fragment. The distribution of its size can itself be computed and so forth, renormalizing again, until the whole available fragments population has been visited. In this way, not only the visiting order of the different fragments can be understood but also their sizes. The purpose of this section is to describe the statistical structure of the size-biased permutation of the fragments' sizes as those obtained while avoiding the ones previously encountered in a sampling process from Dirichlet partition.

• *The RAM structure of size-biased permutation.* Let $\boldsymbol{S}_n := (S_1, \ldots, S_n)$ be the random partition of the interval $[0, 1]$ considered here. Let L_1 be the length of the first randomly chosen fragment $M_1 := M$, so with $L_1 := S_{M_1}$ and $\boldsymbol{P}(M_1 = m_1 \mid \boldsymbol{S}_n) = S_{m_1}$. One may check that $L_1 \stackrel{d}{\sim}$ beta$(1+\theta, (n-1)\theta)$. A standard problem is to iterate the size-biased picking procedure, by avoiding the fragments already encountered: by doing so, a size-biased permutation (SBP) of the fragments is obtained. It turns out that SBP$(\boldsymbol{S}_n)$ has a residual allocation model (RAM) structure.

In the first step of this size-biased picking procedure indeed, $\boldsymbol{S}_n =: \boldsymbol{S}_n^{(0)} \to (L_1, S_1, \ldots, S_{M_1-1}, S_{M_1+1}, \ldots, S_n)$ which may be written as $\boldsymbol{S}_n \to (L_1, (1-L_1)\boldsymbol{S}_{n-1}^{(1)})$, with $\boldsymbol{S}_{n-1}^{(1)} := (S_1^{(1)}, \ldots, S_{M_1-1}^{(1)}, S_{M_1+1}^{(1)}, \ldots, S_n^{(1)})$ a new random partition of the unit interval into $n-1$ random fragments.

Given $L_1 \overset{d}{\sim} \text{beta}(1+\theta, (n-1)\theta)$, the conditional joint distribution of the remaining components of $\boldsymbol{S}_n$ is the same as that of $(1-L_1)\boldsymbol{S}_{n-1}^{(1)}$ where the $(n-1)-$vector $\boldsymbol{S}_{n-1}^{(1)} \overset{d}{\sim} D_{n-1}(\theta)$ has the distribution of a Dirichlet random partition into $n-1$ fragments (see Kingman (1993), Chapter 9). Furthermore, $\boldsymbol{S}_{n-1}^{(1)}$ is independent of $1-L_1$. Pick next at random an interval in $\boldsymbol{S}_{n-1}^{(1)}$ and call V_2 its length, now with distribution $\text{beta}(1+\theta, (n-2)\theta)$, and iterate until all fragments have been exhausted.

With $V_1 := L_1$, the length of the second fragment by avoiding the first reads $L_2 = (1-V_1)V_2$. Iterating, the final SBP of $\boldsymbol{S}_n$ is $\boldsymbol{L}_n := (L_1, \ldots, L_n)$ and we shall put $\boldsymbol{L}_n = \text{SBP}(\boldsymbol{S}_n)$. From this construction, we easily get

LEMMA 2.1. *Let $\boldsymbol{L}_n = SBP(\boldsymbol{S}_n)$. If $(V_1, \ldots, V_{n-1})$ is an independent sample with distribution $V_k \overset{d}{\sim} beta(1+\theta, (n-k)\theta)$, $k = 1, \ldots, n-1$, then,*

$$L_k = \prod_{i=1}^{k-1}(1-V_i)V_k, \qquad k = 1, \ldots, n-1 \tag{2.3}$$

$$L_n = 1 - \sum_{k=1}^{n-1} L_k = \prod_{k=1}^{n-1}(1-V_i) \tag{2.4}$$

is the RAM representation of the size-biased permutation $\boldsymbol{L}_n$ of $\boldsymbol{S}_n$.

Note that $\overline{V}_i := 1 - V_i \overset{d}{\sim} \text{beta}((n-i)\theta, 1+\theta)$ and that V_n should be set to 1. These are well-known construction and properties; see Kingman ((1993), Chapter 9, Section 6) and Donnelly (1986, 1991).

This RAM representation allows to compute the joint distribution of the size-biased permutation $\boldsymbol{L}_n$ of $\boldsymbol{S}_n$. We shall say in the sequel that, if $\boldsymbol{L}_n = \text{SBP}(\boldsymbol{S}_n)$, then $\boldsymbol{L}_n \overset{d}{\sim} SBD_n(\theta)$ assuming that $\boldsymbol{S}_n \overset{d}{\sim} D_n(\theta)$. Before addressing this problem, we shall first consider the order in which fragments are visited.

• *The visiting order of the fragments in the SBP process.* For any permutation $\{m_1, \ldots, m_n\}$ of $\{1, \ldots, n\}$, with $M_1', \ldots, M_k'$, $k = 1, \ldots, n$, the first k *distinct* fragments numbers which have been visited in the SBP sampling process, we have

$$\boldsymbol{P}(M_1' = m_1, \ldots, M_k' = m_k \mid \boldsymbol{S}_n) = \prod_{i=1}^{k-1} \frac{S_{m_i}}{1 - \sum_{l=1}^{i} S_{m_l}} S_{m_k}, \tag{2.5}$$

so that

$$\boldsymbol{P}(M_k' = m_k \mid \boldsymbol{S}_n, M_1' = m_1, \ldots, M_{k-1}' = m_{k-1}) = \frac{S_{m_k}}{1 - \sum_{l=1}^{k-1} S_{m_l}}. \tag{2.6}$$

As a result,

$$\boldsymbol{P}(M_k' = m \mid \boldsymbol{S}_n) = S_m \sum_{(m_1 \neq \cdots \neq m_{k-1}) \neq m} \prod_{i=1}^{k-1} \frac{S_{m_i}}{1 - \sum_{l=1}^{i} S_{m_l}} \tag{2.7}$$

is the conditional (given $\boldsymbol{S}_n$) probability that the k-th visited fragment is fragment number m from $D_n(\theta)$.

LEMMA 2.2. *Given $M'_1 = m_1, \ldots, M'_{k-1} = m_{k-1}$, M'_k is uniformly distributed on the set $m \in \{1, \ldots, n\} \backslash \{m_1 \neq \cdots \neq m_{k-1}\}$. The joint probability distribution of $M'_1, \ldots, M'_k$, $k = 1, \ldots, n$, is Bose-Einstein distribution*

$$\boldsymbol{P}_\theta(M'_1 = m_1, \ldots, M'_k = m_k) = \frac{1}{\prod_{i=0}^{k-1}(n-i)}, \tag{2.8}$$

with $m_1 \neq \cdots \neq m_k \in \{1, \ldots, n\}$.

PROOF. Although this result is immediate by symmetry, we shall supply a short proof of it. From equation (2.6), the function $\boldsymbol{S}_n \to \frac{S_{m_k}}{1-\sum_{l=1}^{k-1} S_{m_l}} = \frac{S_{m_k}}{\sum_{m \neq \{m_1, \ldots, m_{k-1}\}} S_m}$ is homogeneous with degree 0. Applying (ii) of Theorem 1 p. 471 of Huillet and Martinez (2003), with $\{T_1, \ldots, T_n\}$ iid gamma(θ) distributed random variables, we get

$$\begin{aligned} \boldsymbol{P}_\theta(M'_k = m_k \mid M'_1 = m_1, \ldots, M'_{k-1} = m_{k-1}) &= \boldsymbol{E}\left[\frac{S_{m_k}}{\sum_{m \neq \{m_1, \ldots, m_{k-1}\}} S_m}\right] \\ &= \boldsymbol{E}\left[\frac{T_k}{\sum_{l=k}^n T_l}\right] = \frac{1}{n-k+1}. \end{aligned}$$

The Bose-Einstein distribution of $M'_1, \ldots, M'_k$ results from Bayes formula. □

• *The joint distribution of the size-biased permutation.* The SBP of $\boldsymbol{S}_n$ is $\boldsymbol{L}_n$ with $\boldsymbol{L}_n \overset{d}{\sim} SBD_n(\theta)$ and $\boldsymbol{S}_n \overset{d}{\sim} D_n(\theta)$. First, we have

$$(L_1, \ldots, L_n) = (S_{M'_1}, \ldots, S_{M'_n}), \tag{2.9}$$

and consequently

$$\boldsymbol{P}(L_1 = S_{m_1}, \ldots, L_n = S_{m_n} \mid \boldsymbol{S}_n) = \prod_{k=1}^{n-1} \frac{S_{m_k}}{1 - \sum_{l=1}^k S_{m_l}} S_{m_n}. \tag{2.10}$$

Consider now the joint moment function of the random size-biased permutation $\boldsymbol{L}_n = (L_1, \ldots, L_n)$. In Barrera *et al.* (2005), using the RAM representation of $\boldsymbol{L}_n$, the following result was proven.

THEOREM 2.1. *The joint moment function of the SBP $\boldsymbol{L}_n = (L_1, \ldots, L_n) \overset{d}{\sim} SBD_n(\theta)$ reads*

$$\begin{aligned} \boldsymbol{E}\left[\prod_{k=1}^n L_k^{q_k}\right] &= \sum_{\{m_1 \neq \cdots \neq m_n\}} \boldsymbol{E}\left[\prod_{k=1}^{n-1} \frac{S_{m_k}^{q_k+1}}{1 - \sum_{l=1}^k S_{m_l}} S_{m_n}^{q_n+1}\right] \\ &= \prod_{k=1}^{n-1} \left\{ \frac{\Gamma(1 + (n-k+1)\theta)}{\Gamma(1+\theta)\Gamma((n-k)\theta)} \right. \\ &\qquad \left. \times \frac{\Gamma(1+\theta+q_k)\Gamma((n-k)\theta + q_{k+1} + \cdots + q_n)}{\Gamma(1 + (n-k+1)\theta + q_k + \cdots + q_n)} \right\}. \end{aligned} \tag{2.11}$$

• *One-dimensional marginals.* From Theorem 2.1, we get the one-dimensional law of the L_ks, $k = 1, \ldots, n$. Furthermore, one may check that the L_ks are arranged in stochastically decreasing order (denoted by $\succeq$). More precisely (see Barrera *et al.* (2005))

COROLLARY 2.1. (i) *The law of L_k, for $k = 1, \ldots, n$, is characterized by*

$$\boldsymbol{E}[L_k^q] = \prod_{i=1}^{k-1} \boldsymbol{E}[\overline{V}_i^q]\boldsymbol{E}[V_k^q] = \prod_{i=1}^{k-1} \frac{\Gamma((n-i)\theta+q)\Gamma((n-i+1)\theta+1)}{\Gamma((n-i)\theta)\Gamma((n-i+1)\theta+1+q)} \times \frac{\Gamma(1+\theta+q)\Gamma(1+(n-k+1)\theta)}{\Gamma(1+\theta)\Gamma(1+(n-k+1)\theta+q)}. \tag{2.12}$$

(ii) *Let $B_{(n-k+1)\theta,1} \overset{d}{\sim}$ beta$((n-k+1)\theta, 1)$. Then,*

$$L_k \overset{d}{=} B_{(n-k+1)\theta,1} \cdot L_{k-1}, \qquad k = 2, \ldots, n, \tag{2.13}$$

where pairs $B_{(n-k+1)\theta,1}$ and L_{k-1} are mutually independent for $k = 2, \ldots, n$.

(iii) $L_1 \succeq \cdots \succeq L_k \succeq \cdots \succeq L_n$.

The Kingman limit

Consider the limit $n \uparrow \infty$, $\theta \downarrow 0$ while $n\theta = \gamma > 0$. Such an asymptotic was first considered by Kingman (1975); we shall "star" the results (as in $\underset{*}{\overset{d}{\to}}$) when referring to such an asymptotic. As noted by Kingman, $\boldsymbol{S}_n \overset{d}{\sim} D_n(\theta)$ itself has no non-degenerate limit.

When $k = o(n)$, recalling $V_k \overset{d}{\sim}$ beta$(1+\theta, (n-k)\theta)$, we have $V_k \underset{*}{\overset{d}{\to}} V_k^* \overset{d}{\sim}$ beta$(1, \gamma)$ and the $SBD_n(\theta)$ distribution converges weakly from equations (2.3), (2.4) to a Griffiths-Engen-McCloskey or $GEM(\gamma)$ distribution.

Namely, $(L_1, \ldots, L_n) \underset{*}{\overset{d}{\to}} (L_1^*, \ldots, L_k^*, \ldots) =: \boldsymbol{L}^*$ where

$$L_k^* = \prod_{i=1}^{k-1} \overline{V}_i^* V_k^*, \qquad k \geq 1. \tag{2.14}$$

Here $(V_k^*, k \geq 1)$ are iid with common law $V_1^* \overset{d}{\sim}$ beta$(1, \gamma)$ and $\overline{V}_1^* := 1 - V_1^* \overset{d}{\sim}$ beta$(\gamma, 1)$. Note that $L_1^* \succeq \cdots \succeq L_k^* \succeq \cdots$, and that $\boldsymbol{L}^*$ is invariant under size-biased permutation. In the Kingman limit, $(S_{(m)}, m = 1, \ldots, n)$ converges in law to a Poisson-Dirichlet distribution $(L_{(k)}^*, k \geq 1) \overset{d}{\sim} PD(\gamma)$ with $L_{(1)}^* > \cdots > L_{(k)}^* > \cdots$. The size-biased permutation of $(L_{(k)}^*, k \geq 1)$ is $(L_k^*, k \geq 1) \overset{d}{\sim} GEM(\gamma)$ (see Kingman (1993), Chapter 9).

The model (2.14) generates a random countable partition of the unit interval, with many fundamental invariance properties (for a review of these results and applications to Computer Science, Combinatorial Structures, Physics, Biology ..., see Tavaré and

Ewens (1997) and the references therein for example; this model and related ones are also fundamental in Probability Theory; see Pitman (1996, 1999, 2002) and Pitman and Yor (1997).

3. Dirichlet partitions: sampling formulae for unordered and ordered sequences

Ewens' sampling formula (ESF) gives the distribution of alleles (different types of genes) in a sample with size k from the Poisson-Dirichlet process $PD(\gamma)$. Alternatively, it can be described in terms of sequential sampling of animals from a countable collection of distinguishable species drawn from $GEM(\gamma)$. It provides the probability of the partition of a sample of k selectively equivalent genes into a number of alleles as population size becomes indefinitely large. Depending on whether the order of appearance of sequentially sampled species matters or not, we are led to the first ESF for unordered sequences or to the Donnelly-Tavaré-Griffiths (DTG) sampling formula for ordered sequences. A third way to describe the sample is to record the number of species in the k-sample with exactly i representatives, $i = 0, \ldots, k$. When doing this while assuming the species have random frequencies following $GEM(\gamma)$ distribution, we are led to a second Ewens Sampling Formula. We recall here the exact expressions of both first, second Ewens and DTG sampling formulae, when sampling is from finite Dirichlet random partitions. These sampling formulae give both ESF and DTG formulae from $GEM(\gamma)$ when passing to the Kingman limit (see Sibuya and Yamato (1995) for further results). Most of the results (and their proofs) presented therein can be found in Huillet (2005).

3.1 *The first Ewens sampling formula for Dirichlet partitions*

We first consider a sampling formula from Dirichlet partitions for which the order in which the consecutive fragments are being discovered in the sampling process is irrelevant.

Let $\boldsymbol{S}_n$ be the above Dirichlet random partition with parameter $\theta > 0$. Let $k > 1$ and $(U_1, \ldots, U_k)$ be k iid uniform random throws on $[0, 1]$. Let then $(M_1, \ldots, M_k)$ be the (conditionally iid) corresponding fragments numbers (or animals' species), with common conditional and unconditional distributions

$$\boldsymbol{P}(M = m \mid \boldsymbol{S}_n) = S_m, \qquad m \in \{1, \ldots, n\} \tag{3.1}$$

$$\boldsymbol{P}_\theta(M = m) := \boldsymbol{E}[\boldsymbol{P}(M = m \mid \boldsymbol{S}_n)] = \boldsymbol{E}S_m = \frac{1}{n}. \tag{3.2}$$

Let $\mathcal{B}_{n,k}(m) = \sum_{l=1}^{k} \boldsymbol{I}(M_l = m)$ count the random number of occurrences of fragment m in the k-sample and $P_{n,k} := \sum_{m=1}^{n} \boldsymbol{I}(\mathcal{B}_{n,k}(m) > 0)$ count the number of distinct fragments which have been visited in the k-sampling process. We let $(\theta)_k := \theta(\theta+1)(\theta+k-1)$. In Huillet (2005) we obtained

THEOREM 3.1. (i) *We have*

$$\begin{aligned} &\boldsymbol{P}_\theta(\mathcal{B}_{n,k}(1) = k_1, \ldots, \mathcal{B}_{n,k}(p) = k_p; P_{n,k} = p) \\ &\quad = \binom{n}{p} \frac{k!}{\prod_{q=1}^{p} k_q!} \frac{1}{(n\theta)_k} \prod_{q=1}^{p} (\theta)_{k_q}. \end{aligned} \tag{3.3}$$

(ii) *With*

$$B_{k,p}(x_1, x_2, \ldots) := \sum_{\substack{a_i \geq 0: \sum_{i=1}^{k} i a_i = k; \\ \sum_{i=1}^{k} a_i = p}} \frac{k!}{\prod_{i=1}^{k} i!^{a_i} a_i!} \prod_{i=1}^{k} x_i^{a_i}$$

the Bell polynomials,

$$\boldsymbol{P}_\theta(P_{n,k} = p) = \frac{n!}{(n-p)!} \frac{1}{(n\theta)_k} B_{k,p}((\theta)_1, (\theta)_2, \ldots). \tag{3.4}$$

(iii) *We have*

$$\begin{aligned} &\boldsymbol{P}_\theta(\mathcal{B}_{n,k}(1) = k_1, \ldots, \mathcal{B}_{n,k}(p) = k_p \mid P_{n,k} = p) \\ &\quad = \frac{k!}{p!} \frac{1}{B_{k,p}((\theta)_1, (\theta)_2, \ldots)} \prod_{q=1}^{p} \frac{(\theta)_{k_q}}{k_q!}. \end{aligned} \tag{3.5}$$

Note in particular that if $p = k = 2$, with $k_1 = k_2 = 1$, equation (3.3) gives

$$\boldsymbol{P}_\theta(\mathcal{B}_{n,2}(1) = 1, \mathcal{B}_{n,2}(2) = 1; P_{n,2} = 2) = \frac{(n-1)\theta}{n\theta + 1}.$$

This is the probability that the first 2 random throws will visit any 2 distinct fragments. The complementary probability that it does not is thus $1 - \frac{(n-1)\theta}{n\theta+1} = \frac{\theta+1}{n\theta+1}$.

Remark. (the law of succession) We would like to briefly recall a related question raised in Donnelly (1986) and Ewens (1996), concerning the law of succession.

(i) Let the "M_{k+1} is new" denote the event that M_{k+1} is none of the previously visited fragments. One can prove (see Huillet (2005))

$$\begin{aligned} &\boldsymbol{P}_\theta(M_{k+1} \text{ is new} \mid \mathcal{B}_{n,k}(1) = k_1, \ldots, \mathcal{B}_{n,k}(p) = k_p; P_{n,k} = p) \\ &\quad = \frac{(n-p)\theta}{n\theta + k}, \end{aligned} \tag{3.6}$$

which is independent of cell occupancies $k_1, \ldots, k_p$ but depends on the number p of distinct fragments already visited by the k-sample.

(ii) Similarly, let the event "$M_{k+1} \in$ species seen k_r times" denote the fact that the $(k+1)$-th sample is one from the previously encountered fragment already visited k_r times. We easily get

$$\begin{aligned} &\boldsymbol{P}_\theta(M_{k+1} \in \text{species seen } k_r \text{ times} \mid \mathcal{B}_{n,k}(1) = k_1, \ldots, \mathcal{B}_{n,k}(p) = k_p; P_{n,k} = p) \\ &\quad = \frac{\theta + k_r}{n\theta + k}, \end{aligned} \tag{3.7}$$

which is independent of occupancy numbers k_q, $q \in \{1, \ldots, p\} \backslash \{r\}$ and also of the number p of distinct observations.

Remark. (number of distinct observations) From equations (3.6) and (3.7), we also have the transition probabilities

$$\boldsymbol{P}_\theta(P_{n,k+1} = p+1 \mid P_{n,k} = p) = \frac{(n-p)\theta}{n\theta + k}$$

$$\boldsymbol{P}_\theta(P_{n,k+1} = p \mid P_{n,k} = p) = \frac{\sum_{r=1}^{p}(\theta + k_r)}{n\theta + k} = \frac{p\theta + k}{n\theta + k}.$$

Next,

$$\boldsymbol{P}_\theta(P_{n,k+1} = p) = \frac{(n-p+1)\theta}{n\theta + k}\boldsymbol{P}_\theta(P_{n,k} = p-1) + \frac{p\theta + k}{n\theta + k}\boldsymbol{P}_\theta(P_{n,k} = p).$$

Using equation (3.4), we obtain the following triangular recurrence for Bell polynomials $B_{k,p}((\theta)_1, (\theta)_2, \ldots)$,

$$B_{k+1,p}((\theta)_1, (\theta)_2, \ldots) = \theta B_{k,p-1}((\theta)_1, (\theta)_2, \ldots) + (p\theta + k)B_{k,p}((\theta)_1, (\theta)_2, \ldots).$$

These should be considered with boundary conditions

$$B_{k,0}((\theta)_1, (\theta)_2, \ldots) = B_{0,p}((\theta)_1, (\theta)_2, \ldots) = 0,$$

except for $B_{0,0}((\theta)_1, (\theta)_2, \ldots) := 1$.

This leads in particular to $B_{k,1}((\theta)_1, (\theta)_2, \ldots) = (\theta)_k$, $k \geq 1$ and to

$$\boldsymbol{P}_\theta(P_{n,k} = 1) = \frac{n(\theta)_k}{(n\theta)_k}.$$

The Kingman limit

Consider the situation where $n \uparrow \infty$, $\theta \downarrow 0$ while $n\theta = \gamma > 0$ in equations (3.3), (3.4) of Theorem 3.1. Proceeding in this way, we recover the first Ewens sampling formula (1972):

COROLLARY 3.1. (i) *In the Kingman limit,*

$$\boldsymbol{P}_\theta(\mathcal{B}_{n,k}(1) = k_1, \ldots, \mathcal{B}_{n,k}(p) = k_p; P_{n,k} = p)$$

converges to

$$\boldsymbol{P}^*_\gamma(\mathcal{B}_k(1) = k_1, \ldots, \mathcal{B}_k(p) = k_p; P_k = p) = \frac{k!}{p!}\frac{\gamma^p}{(\gamma)_k \prod_{q=1}^{p} k_q}. \tag{3.8}$$

(ii) *With $s_{k,p}$ the absolute value of the first kind Stirling numbers, we get*

$$\boldsymbol{P}^*_\gamma(P_k = p) = \frac{\gamma^p s_{k,p}}{(\gamma)_k}, \qquad p = 1, \ldots, k \tag{3.9}$$

and

$$\boldsymbol{P}^*_\gamma(\mathcal{B}_k(1) = k_1, \ldots, \mathcal{B}_k(p) = k_p \mid P_k = p) = \frac{k!}{p!}\frac{1}{s_{k,p}\prod_{q=1}^{p} k_q}. \tag{3.10}$$

Remark. (the law of succession) In the Kingman limit, the probabilities displayed in examples (3.6) and (3.7) converge respectively to

$$\frac{\gamma}{\gamma + k} \quad \text{and} \quad \frac{k_r}{\gamma + k}. \tag{3.11}$$

3.2 *The second Ewens formula for Dirichlet populations*

Let now $\mathcal{A}_{n,k}(i)$, $i \in \{0,\dots,k\}$ count the number of fragments in the k-sample with i representatives, that is

$$\mathcal{A}_{n,k}(i) =^{\#} \{m \in \{1,\dots,n\} : \mathcal{B}_{n,k}(m) = i\} = \sum_{m=1}^{n} \boldsymbol{I}(\mathcal{B}_{n,k}(m) = i). \tag{3.12}$$

Then $\sum_{i=0}^{k} \mathcal{A}_{n,k}(i) = n$, $\sum_{i=1}^{k} \mathcal{A}_{n,k}(i) = p$ is the number of fragments visited by the k-sample and $\mathcal{A}_{n,k}(0)$ the number of unvisited ones. Note that $\sum_{i=1}^{k} i\mathcal{A}_{n,k}(i) = k$ is the sample size.

The vector $(\mathcal{A}_{n,k}(1),\dots,\mathcal{A}_{n,k}(k))$ is called the fragments vector count or the species vector count in biology, see Ewens (1990). In this case, see Huillet (2005), we have

THEOREM 3.2. (i) *For any* $a_i \geq 0$, $i = 1,\dots,k$ *satisfying* $\sum_{i=1}^{k} ia_i = k$ *and* $\sum_{i=1}^{k} a_i = p$, *we have*

$$\begin{aligned} &\boldsymbol{P}_\theta(\mathcal{A}_{n,k}(1) = a_1,\dots,\mathcal{A}_{n,k}(k) = a_k; P_{n,k} = p) \\ &\quad = \frac{n!}{(n-p)!} \frac{k!}{\prod_{i=1}^{k} i!^{a_i} a_i!} \frac{1}{(n\theta)_k} \prod_{i=1}^{k} (\theta)_i^{a_i}. \end{aligned} \tag{3.13}$$

(ii) *With* $B_{k,p}(x_1,x_2,\dots)$, *the Bell polynomials, we have*

$$\boldsymbol{P}_\theta(P_{n,k} = p) = \frac{n!}{(n-p)!} \frac{\Gamma(n\theta)}{\Gamma(n\theta+k)} B_{k,p}((\theta)_1,(\theta)_2,\dots). \tag{3.14}$$

(iii) *We have*

$$\begin{aligned} &\boldsymbol{P}_\theta(\mathcal{A}_{n,k}(1) = a_1,\dots,\mathcal{A}_{n,k}(k) = a_k \mid P_{n,k} = p) \\ &\quad = \frac{k!}{B_{k,p}((\theta)_1,(\theta)_2,\dots)} \prod_{i=1}^{k} \frac{(\theta)_i^{a_i}}{i!^{a_i} a_i!}. \end{aligned} \tag{3.15}$$

The Kingman limit

Consider the limit $n \uparrow \infty$, $\theta \downarrow 0$ while $n\theta = \gamma > 0$. We recover the celebrated Ewens Sampling Formula (1972).

COROLLARY 3.2. (i) *In the Kingman limit, the probability displayed in* (3.13) *converges to*

$$\boldsymbol{P}^*_\gamma(\mathcal{A}_k(1) = a_1,\dots,\mathcal{A}_k(k) = a_k; P_k = p) = \frac{k!\gamma^p}{(\gamma)_k \prod_{i=1}^{k} i^{a_i} a_i!}. \tag{3.16}$$

(ii) *With* $s_{k,p}$ *the absolute value of the first kind Stirling numbers,*

$$\boldsymbol{P}^*_\gamma(P_k = p) = \frac{\gamma^p s_{k,p}}{(\gamma)_k}, \qquad p = 1,\dots,k \tag{3.17}$$

and

$$\boldsymbol{P}^*_\gamma(\mathcal{A}_k(1) = a_1,\dots,\mathcal{A}_k(p) = a_k \mid P_k = p) = \frac{k!}{s_{k,p} \prod_{i=1}^{k} i^{a_i} a_i!}. \tag{3.18}$$

3.3 *Donnelly-Tavaré-Griffiths sampling formula for the Dirichlet partition*

We now consider sampling formulae from Dirichlet partitions for which the order in which the consecutive fragments are being discovered in the sampling process matters.

Consider a k-sample and let $m_1 \neq m_2 \neq \cdots \neq m_p$ denote the ordered number of the first, second, ..., the p-th distinct animals sampled from $\boldsymbol{S}_n$ when only $P_{n,k} = p$ distinct fragments were visited. Let $\mathcal{C}_{n,k}(q)$, $q = 1, \ldots, p$ be the number of animals of the q-th species to appear. Using Theorem 2.1, we can prove (see Huillet (2005))

THEOREM 3.3. *For any* $c_q \geq 1$, $q = 1, \ldots, p$ *satisfying* $\sum_1^p c_q = k$ *and any* $p = 1, \ldots, n \wedge k$,

$$
\begin{aligned}
(3.19)\quad & \boldsymbol{P}_\theta(\mathcal{C}_{n,k}(1) = c_1, \ldots, \mathcal{C}_{n,k}(p) = c_p; P_{n,k} = p) \\
&= \frac{(k-1)!}{\prod_{q=1}^{p-1}(k - \sum_1^q c_i)} \frac{\Gamma(1 + (n-p+1)\theta)\Gamma(\theta + c_p)}{\Gamma(1+\theta)\Gamma((n-p+1)\theta + c_p)\Gamma(c_p)} \\
&\quad \times \prod_{q=1}^{p-1} \left\{ \frac{\Gamma(1 + (n-q+1)\theta)\Gamma(\theta + c_q)}{\Gamma(1+\theta)\Gamma((n-q)\theta)\Gamma(c_q)} \frac{\Gamma((n-q)\theta + c_{q+1} + \cdots + c_p)}{\Gamma((n-q+1)\theta + c_q + \cdots + c_p)} \right\}.
\end{aligned}
$$

Remark. (the law of succession) (i) Consider equation (3.19) and, with $m_r \in \{m_1, \ldots, m_p\}$, let

$$
\boldsymbol{P}(M_{k+1} = m_r \mid \mathcal{C}_{n,k}(1) = c_1, \ldots, \mathcal{C}_{n,k}(p) = c_p; P_{n,k} = p)
$$

be the conditional probability that the $(k+1)$-th sample is one from the previously encountered species already visited c_r times. One can show, see Huillet (2005), that, as for the ESF,

$$
\begin{aligned}
(3.20)\quad & \boldsymbol{P}_\theta(M_{k+1} \in \text{ species seen } c_r \text{ times} \mid \mathcal{C}_{n,k}(1) = c_1, \ldots, \mathcal{C}_{n,k}(p) = c_p; P_{n,k} = p) \\
&= \frac{\theta + c_r}{n\theta + k},
\end{aligned}
$$

which is again independent of c_q, $q \in \{1, \ldots, p\} \backslash \{r\}$ and also of p.

(ii) Summing over $r = 1, \ldots, p$, the conditional probability that $M_{k+1} \in \{$any one of the species previously seen$\}$ is thus $\sum_{r=1}^p \frac{\theta + c_r}{n\theta + k} = \frac{p\theta + k}{n\theta + k}$. Taking its complement to 1, we obtain

$$
\begin{aligned}
(3.21)\quad & \boldsymbol{P}_\theta(M_{k+1} \text{ is new} \mid \mathcal{C}_{n,k}(1) = c_1, \ldots, \mathcal{C}_{n,k}(p) = c_p; P_{n,k} = p) \\
&= \frac{(n-p)\theta}{n\theta + k},
\end{aligned}
$$

which is independent of cell occupancies $c_1, \ldots, c_p$ but depends on the number p of distinct fragments already visited by the k-sample.

The Kingman limit

Passing to the Kingman limit in Theorem 3.3 gives the celebrated Donnelly-Tavaré-Griffiths sampling formula given in Donnelly and Tavaré (1986), p. 10 (see Huillet (2005) for details).

COROLLARY 3.3. (i) *In the Kingman limit, the probability* (3.19) *converges to*

$$\boldsymbol{P}^*_\gamma(\mathcal{C}_k(1)=c_1,\dots,\mathcal{C}_k(p)=c_p;P_k=p)=\frac{k!\gamma^p}{(\gamma)_k\prod_{q=1}^p(c_q+\cdots+c_p)}. \tag{3.22}$$

(ii) *With $s_{k,p}$ the absolute value of the first kind Stirling numbers,*

$$\boldsymbol{P}^*_\gamma(P_k=p)=\frac{\gamma^p s_{k,p}}{(\gamma)_k},\qquad p=1,\dots,k \tag{3.23}$$

and

$$\boldsymbol{P}^*_\gamma(\mathcal{C}_k(1)=c_1,\dots,\mathcal{C}_k(p)=c_p\mid P_k=p)=\frac{k!}{s_{k,p}\prod_{q=1}^p(c_q+\cdots+c_p)}. \tag{3.24}$$

4. Intervals between new sampled species from Dirichlet partition

In this section, we shall consider the following sampling problem from Dirichlet partition. From the knowledge of the order in which new fragments are being discovered, what can be said about the random number of samples separating the discovery of consecutive new fragments until exhaustion of the list? We shall first consider the case of a fixed sample size k and then the case of an unlimited size.

Assume there are $P_{n,k}=p\le n\wedge k:=k_n$ distinct species visited by the k-sample drawn from $\boldsymbol{S}_n\overset{d}{\sim}D_n(\theta)$, with $k\ge 2$. Let $m_1\ne\cdots\ne m_p$ be their fragments numbers in order of their appearance. Let $\mathcal{D}_{n,k}(r)$, $r=1,\dots,p-1$ be the sample size between the discovery of the r-th and the $(r+1)$-th new fragments. Let $\chi_1=1$ and $\chi_l=\prod_{k=1}^{l-1}\boldsymbol{I}(M_l\ne M_k)$, $l=2,\dots,k$. This binary sequence takes the value 1 each time a new species is visited by the k-sample; otherwise its value is 0. Let $d_r\ge 1$, $r=1,\dots,p-1$, $d_p:=k-\sum_{r=1}^{p-1}d_r\ge 1$ and $\bar{d}_r=\sum_{q=1}^r d_q$, $r=1,\dots,p$, $\bar{d}_0:=0$. The following two events are identical

$$\text{“}\mathcal{D}_{n,k}(1)=d_1,\dots,\mathcal{D}_{n,k}(p-1)=d_{p-1};P_{n,k}=p\text{”}$$

and

$$\begin{aligned}\text{“}&\chi_1=1,\chi_2=\cdots=\chi_{d_1}=0,\chi_{d_1+1}=1,\chi_{d_1+2}=\cdots\\&=\chi_{\bar{d}_2}=0,\dots,\chi_{\bar{d}_{p-1}+1}=1,\chi_{\bar{d}_{p-1}+2}=\cdots=\chi_{\bar{d}_p}=0\text{”}.\end{aligned}$$

Next, from the above law of succession displayed in (3.21), with $b_1,\dots,b_l\in\{0,1\}^l$, we get

$$\begin{aligned}\boldsymbol{P}_\theta(\chi_{l+1}=1\mid\chi_1=b_1,\dots,\chi_l=b_l)&=\frac{(n-\sum_{q=1}^l b_q)\theta}{n\theta+l}\\ \boldsymbol{P}_\theta(\chi_{l+1}=0\mid\chi_1=b_1,\dots,\chi_l=b_l)&=\frac{l+\theta\cdot\sum_{q=1}^l b_q}{n\theta+l}.\end{aligned}$$

Using this, we obtain

THEOREM 4.1. *Let $p\le k_n$. With $d_r\ge 1$, $r=1,\dots,p-1$, $d_p:=k-\sum_{r=1}^{p-1}d_r\ge 1$ and $\bar{d}_r=\sum_{q=1}^r d_q$, $r=1,\dots,p$, $\bar{d}_0:=0$, we get*

(i)

$$\boldsymbol{P}_\theta(\mathcal{D}_{n,k}(1)=d_1,\ldots,\mathcal{D}_{n,k}(p-1)=d_{p-1};P_{n,k}=p) = \frac{\theta^p n!}{(n-p)!(n\theta)_k}\prod_{r=1}^{p}(1+r\theta+\overline{d}_{r-1})_{d_r-1}. \tag{4.1}$$

(ii)

$$\boldsymbol{P}_\theta(\mathcal{D}_{n,k}(1)=d_1,\ldots,\mathcal{D}_{n,k}(p-1)=d_{p-1}\mid P_{n,k}=p) = \frac{\theta^p}{B_{k,p}((\theta)_1,(\theta)_2,\ldots)}\prod_{r=1}^{p}(1+r\theta+\overline{d}_{r-1})_{d_r-1}. \tag{4.2}$$

PROOF. (i) We have:

$$\begin{aligned}
&\boldsymbol{P}_\theta(\mathcal{D}_{n,k}(1)=d_1,\ldots,\mathcal{D}_{n,k}(p-1)=d_{p-1};P_{n,k}=p)\\
&=\prod_{r=0}^{p-1}\frac{(n-r)\theta}{\overline{d}_r+n\theta}\cdot\prod_{r=1}^{p}\prod_{l=\overline{d}_{r-1}+1}^{\overline{d}_r-1}\frac{l+r\theta}{l+n\theta}\\
&=\frac{\prod_{r=0}^{p-1}[(n-r)\theta]}{\prod_{l=0}^{k-1}(l+n\theta)}\prod_{r=1}^{p}\prod_{l=\overline{d}_{r-1}+1}^{\overline{d}_r-1}(l+r\theta)\\
&=\frac{n!}{(n-p)!}\frac{\theta^p}{(n\theta)_k}\prod_{r=1}^{p}(1+r\theta+\overline{d}_{r-1})_{d_r-1},
\end{aligned}$$

(ii) follows from the expression (3.14) of $\boldsymbol{P}_\theta(P_{n,k}=p)$. □

Remark. (Waring distribution) Let us recall here some elementary facts on Waring distributions in the generalized hypergeometric of type $B3$ class. A discrete random variable $W\in\{0,1,\ldots\}$ has Waring distribution $W(b,a)$ with parameters $b>a>0$ if $P(W=x)=(b-a)\frac{(a)_x}{(b)_{x+1}}$. We have $\boldsymbol{E}(W)=a/(b-a-1)$ if $b-a>1$, $=+\infty$ if not. Regrouping the events "$W\geq n$", a random variable $bW\in\{0,1,\ldots,n\}$ has bounded Waring distribution $bW(n,b,a)$ if $P(bW=x)=(b-a)\frac{(a)_x}{(b)_{x+1}}$, if $x\in\{0,1,\ldots,n-1\}$, $P(bW=n)=\frac{(a)_n}{(b)_n}$ if $bW=n$. We obtain $\boldsymbol{E}(bW)=\frac{a}{b-a-1}(1-\frac{(a+1)_n}{(b)_n})$ if $b\neq a+1$ and $\boldsymbol{E}(bW)=(b-1)[\psi'(b+n)-\psi'(b)]$ if $b=a+1$, where $\psi'(b):=\Gamma'(b)/\Gamma(b)$.

One may check that if $W\overset{d}{\sim}W(b,a)$, then $W\overset{d}{\sim}\text{geom}(\text{beta}(a,b-a))$ where $G\overset{d}{\sim}\text{geom}(p)$ is a geometrically distributed random variable with law $\boldsymbol{P}(G=x)=p^x(1-p)$, $x\geq 0$. Here, p is the random success probability, with $p\overset{d}{\sim}\text{beta}(a,b-a)$, independent of G. From this, we get

$$\begin{aligned}
P(W=x)&=\frac{\Gamma(b)}{\Gamma(a)\Gamma(b-a)}\int_0^1 p^{x+a-1}(1-p)^{b-a}dp\\
&=\frac{\Gamma(b)}{\Gamma(a)\Gamma(b-a)}\frac{\Gamma(a+x)\Gamma(b-a+1)}{\Gamma(b+x+1)}=(b-a)\frac{(a)_x}{(b)_{x+1}}.
\end{aligned}$$

Furthermore, $P(W \geq n) = \frac{\Gamma(b)}{\Gamma(a)\Gamma(b-a)} \int_0^1 p^{n+a-1}(1-p)^{b-a-1}dp = \frac{(a)_n}{(b)_n}$. Let bG have bounded geometric distribution such that $P(bG = x) = p^x(1-p)$, if $x \in \{0, 1, \ldots, n-1\}$, $P(bG = n) = p^n$ if $bG = n$. Using $\boldsymbol{E}(G) = p/(1-p)$ and $\boldsymbol{E}(bG) = \frac{p}{1-p} - \frac{p^{n+1}}{1-p}$, the claims on $\boldsymbol{E}(W)$ and $\boldsymbol{E}(bW)$ follow from randomization of p.

Now, we are ready for presenting subsequent results. As a consequence of Theorem 4.1, we have

THEOREM 4.2. *Let* $1 \leq r < k_n$.
(i) *With* $d_1, \ldots, d_r \geq 1$, $\bar{d}_r < k$, *we have*

$$\begin{aligned} (4.3) \qquad & \boldsymbol{P}_\theta(\mathcal{D}_{n,k}(1) = d_1, \ldots, \mathcal{D}_{n,k}(r) = d_r) \\ & = \frac{(n-1)!}{(n-r-1)!} \frac{\theta^r}{(1+n\theta)_{\bar{d}_r}} \prod_{q=1}^{r} (1 + q\theta + \bar{d}_{q-1})_{d_q - 1}. \end{aligned}$$

In particular, if $r = 1$

$$\begin{aligned} (4.4) \qquad \boldsymbol{P}_\theta(\mathcal{D}_{n,k}(1) = d_1) &= \theta(n-1)\frac{(1+\theta)_{d_1-1}}{(1+n\theta)_{d_1}}, \qquad d_1 \in \{1, \ldots, k_n - 1\} \\ &= \frac{(1+\theta)_{d_1-1}}{(1+n\theta)_{d_1-1}}, \qquad \text{if} \quad d_1 = k_n, \end{aligned}$$

and $\mathcal{D}_{n,k}(1) - 1$ *has bounded Waring distribution* $bW(k_n - 1, 1 + n\theta, 1 + \theta)$. *Consequently,*

$$(4.5) \qquad \boldsymbol{E}_\theta[\mathcal{D}_{n,k}(1)] = 1 + \frac{1+\theta}{(n-1)\theta - 1}\left[1 - \frac{(2+\theta)_{k_n-1}}{(1+n\theta)_{k_n-1}}\right].$$

(ii) *Let* $1 \leq r < k_n$. *With* $d_1, \ldots, d_r \geq 1$, $\bar{d}_{r-1} =: d \geq r - 1$, $\bar{d}_r = d + d_r < k$, *we have*

$$\begin{aligned} (4.6) \qquad & \boldsymbol{P}_\theta(\mathcal{D}_{n,k}(r) = d_r \mid \mathcal{D}_{n,k}(1) = d_1, \ldots, \mathcal{D}_{n,k}(r-1) = d_{r-1}) \\ & = \theta(n-r)\frac{(1+r\theta+d)_{d_r-1}}{(1+n\theta+d)_{d_r}}, \qquad \text{if} \quad d_r \in \{1, \ldots, k_n - d - 1\} \\ & = \frac{(1+\theta+d)_{d_r-1}}{(1+n\theta+d)_{d_r-1}}, \qquad \text{if} \quad d_r = k_n - d, \end{aligned}$$

showing that, given $\overline{\mathcal{D}}_{n,k}(r-1) := \sum_{q=1}^{r-1} \mathcal{D}_{n,k}(q) = d$, $\mathcal{D}_{n,k}(r) - 1$ *has bounded Waring distribution* $bW(k_n - d - 1, 1 + n\theta + d, 1 + r\theta + d)$. *Note from the preceding remark that*

$$(4.7) \qquad \boldsymbol{E}_\theta[\overline{\mathcal{D}}_{n,k}(r-1)] = 1 + \frac{1+r\theta+d}{(n-r)\theta - 1}\left[1 - \frac{(2+r\theta+d)_{k_n-d-1}}{(1+n\theta+d)_{k_n-d-1}}\right].$$

(iii) *For* $r = 1, \ldots, k_n - 1$, *the cumulative random variable* $\overline{\mathcal{D}}_{n,k}(r) := \sum_{q=1}^{r} \mathcal{D}_{n,k}(q)$ *has distribution*

$$\begin{aligned} (4.8) \qquad \boldsymbol{P}_\theta(\overline{\mathcal{D}}_{n,k}(r) = d) &= \frac{(n-1)!B_{d,r}((\theta)_1, (\theta)_2, \ldots)}{(n-r-1)!(1+n\theta)_d}, \qquad d = r, \ldots, k_n - 1 \\ &= \frac{n!}{(n-r)!}\frac{B_{d,r}((\theta)_1, (\theta)_2, \ldots)}{(n\theta)_d}, \qquad d = k_n, \end{aligned}$$

where $B_{d,r}((\theta)_1, (\theta)_2, \ldots)$ are Bell polynomials.

PROOF. (i) The following two events coincide

$$\text{“}\mathcal{D}_{n,k}(1) = d_1, \ldots, \mathcal{D}_{n,k}(r) = d_r\text{”}$$

and “$\chi_1 = 1, \chi_2 = \cdots = \chi_{d_1} = 0, \chi_{d_1+1} = 1, \chi_{d_1+2} = \cdots = \chi_{\overline{d}_2} = 0, \ldots, \chi_{\overline{d}_{r-1}+1} = 1, \chi_{\overline{d}_{r-1}+2} = \cdots = \chi_{\overline{d}_r} = 0, \chi_{\overline{d}_r+1} = 1$”. Proceeding as for the proof of Theorem 4.3 gives (i).

In particular,

$$\begin{aligned} \boldsymbol{P}_\theta(\mathcal{D}_{n,k}(1) = d_1) &= \frac{\theta^2 n!}{(n-2)!} \frac{(1+\theta)_{d_1-1}}{(n\theta)_{d_1+1}} \\ &= \frac{(n-1)(\theta)_{d_1}}{(1+n\theta)_{d_1}} = \theta(n-1)\frac{(1+\theta)_{d_1-1}}{(1+n\theta)_{d_1}}, \end{aligned}$$

so that, with $d = 0, \ldots, k_n - 1$

$$\begin{aligned} \boldsymbol{P}_\theta(\mathcal{D}_{n,k}(1) - 1 = d) &= \theta(n-1)\frac{(1+\theta)_d}{(1+n\theta)_{d+1}}, \qquad d = 0, \ldots, k_n - 2 \\ &= \frac{(1+\theta)_d}{(1+n\theta)_d}, \qquad \text{if} \quad d = k_n - 1 \end{aligned}$$

which is a bounded Waring distribution $bW(k_n - 1, 1 + n\theta, 1 + \theta)$. The expected value of $\mathcal{D}_{n,k}(1)$ displayed in (4.5) follows from the preceding remark.

(ii) If $d_r \in \{1, \ldots, k - d - 1\}$, it follows from the law of $\mathcal{D}_{n,k}(1), \ldots, \mathcal{D}_{n,k}(r)$ that

$$\begin{aligned} &\boldsymbol{P}_\theta(\mathcal{D}_{n,k}(r) = d_r \mid \mathcal{D}_{n,k}(1) = d_1, \ldots, \mathcal{D}_{n,k}(r-1) = d_{r-1}) \\ &\quad = \theta(n-r)\frac{(n\theta)_{d+1}}{(n\theta)_{d+d_r+1}}(1 + r\theta + d)_{d_r-1} \\ &\quad = \theta(n-r)\frac{(1+r\theta+d)_{d_r-1}}{(1+n\theta+d)_{d_r}}, \end{aligned}$$

where $d := d_1 + \cdots + d_{r-1}$. This conditional distribution depends on $d_1, \ldots, d_{r-1}$ only through their sum.

From this, one may check that given $\overline{\mathcal{D}}_{n,k}(r-1) = d$, $\mathcal{D}_{n,k}(r) - 1$ has bounded Waring distribution $bW(k_n - d - 1, 1 + n\theta + d, 1 + r\theta + d)$. The expected value of $\overline{\mathcal{D}}_{n,k}(r-1)$ displayed in (4.7) follows from the preceding remark.

(iii) This results from the Ewens sampling formula displayed in Theorem 3.2.

Suppose that $(\mathcal{D}_{n,k}(1), \ldots, \mathcal{D}_{n,k}(p-1); P_{n,k})$ has distribution given by (4.1). Let r and $\overline{d}_r$ be integers ≥ 1 such that $r \leq \overline{d}_r < k_n$. Let $a_1, \ldots, a_{\overline{d}_r} \geq 0$ be an unordered partition of $\overline{d}_r$ for which $\sum_1^{\overline{d}_r} a_i = r$ and $\sum_1^{\overline{d}_r} i a_i = \overline{d}_r$. Then

$$\begin{aligned} &\boldsymbol{P}_\theta(\mathcal{A}_{n,k}^{(r)}(1) = a_1, \ldots, \mathcal{A}_{n,k}^{(r)}(\overline{d}_r) = a_{\overline{d}_r}) \\ &\quad = \sum \boldsymbol{P}_\theta(\mathcal{D}_{n,k}(1) = d_1, \ldots, \mathcal{D}_{n,k}(r) = d_r) \end{aligned}$$

where summation runs over all distinct ordered partitions $d_1, \ldots, d_r$ of $\bar{d}_r$ which give rise to the unordered partition $a_1, \ldots, a_{\bar{d}_r}$ of $\bar{d}_r$. From (4.3), we get

$$\begin{aligned} &\boldsymbol{P}_\theta(\mathcal{A}_{n,k}^{(r)}(1) = a_1, \ldots, \mathcal{A}_{n,k}^{(r)}(\bar{d}_r) = a_{\bar{d}_r}) \\ &= \frac{n!}{(n-r)!} \frac{\bar{d}_r!}{\prod_{i=1}^{\bar{d}_r} i!^{a_i} a_i!} \frac{1}{(n\theta)_{\bar{d}_r}} \prod_{i=1}^{\bar{d}_r} (\theta)_i^{a_i} \quad \text{if} \quad \bar{d}_r = k_n \\ &= \frac{(n-1)!}{(n-r-1)!} \frac{\bar{d}_r!}{\prod_{i=1}^{\bar{d}_r} i!^{a_i} a_i!} \frac{1}{(1+n\theta)_{\bar{d}_r}} \prod_{i=1}^{\bar{d}_r} (\theta)_i^{a_i} \quad \text{if} \quad r \le \bar{d}_r < k_n \end{aligned}$$

with

$$\sum \frac{\bar{d}_r!}{\prod_{i=1}^{\bar{d}_r} i!^{a_i} a_i!} \prod_{i=1}^{\bar{d}_r} (\theta)_i^{a_i} = B_{\bar{d}_r, r}((\theta)_1, (\theta)_2, \ldots).$$

Next, for $d = r, r+1, \ldots, k_n - 1$, we have

$$\boldsymbol{P}_\theta(\bar{\mathcal{D}}_{n,k}(r) = d) = \sum \boldsymbol{P}_\theta(\mathcal{A}_{n,k}^{(r)}(1) = a_1, \ldots, \mathcal{A}_{n,k}^{(r)}(d) = a_d)$$

where summation runs over $a_i \ge 0$ satisfying $\sum_1^d a_i = r$ and $\sum_1^d i a_i = d$. □

Lastly, we shall turn to the related question of computing the law of $\mathcal{D}_n(r)$, the sample size separating consecutive visits to the r-th and the $(r+1)$-th new species, when sampling is from $D_n(\theta)$ with unlimited sample size. For this problem, we obtain

THEOREM 4.3. *Let* $1 \le r < n-1$.

(i) *The law of* $\mathcal{D}_n(r)$ *is given by*

$$\begin{aligned} (4.9) \qquad \boldsymbol{P}_\theta(\mathcal{D}_n(r) > d) = &\sum_{k_1, \ldots, k_r \ge 0 : \sum_1^r k_q = d} \binom{d}{k_1, \ldots, k_r} \\ &\times \prod_{q=1}^{r} \left\{ \frac{\Gamma(1 + (n-q+1)\theta)}{\Gamma(1+\theta)\Gamma((n-q)\theta)} \right. \\ &\left. \times \frac{\Gamma(1+\theta+k_q)\Gamma((n-q)\theta + k_{q+1} + \cdots + k_r)}{\Gamma(1 + (n-q+1)\theta + k_q + \cdots + k_r)} \right\} \end{aligned}$$

and the following stochastic domination property holds

$$\mathcal{D}_n(r) \succeq \mathcal{D}_n(r-1).$$

(ii) *With* $1 \le r < n-1$, *the mean value of* $\mathcal{D}_n(r)$ *reads*

$$\begin{aligned} (4.10) \qquad \boldsymbol{E}_\theta(\mathcal{D}_n(r)) &= \prod_{q=1}^{r} \frac{(n-q+1)\theta}{(n-q)\theta - 1}, \quad \textit{if} \quad (n-r)\theta > 1 \\ &= +\infty, \quad \textit{if not.} \end{aligned}$$

PROOF. (i) Given $\boldsymbol{S}_n$, we have

$$\boldsymbol{P}(\mathcal{D}_n(r) > d \mid M_1' = m_1, \ldots, M_r' = m_r, \boldsymbol{S}_n) = \left(\sum_{q=1}^{r} S_{m_q}\right)^d$$

since, if $\mathcal{D}_n(r) > d$, at least d sample need to fall in the already visited fragments. Averaging over $\boldsymbol{S}_n$ and summing over all realizations $m_1 \neq \cdots \neq m_r$ of $M_1', \ldots, M_r'$, we get

$$\boldsymbol{P}_\theta(\mathcal{D}_n(r) > d) = \boldsymbol{E}\left[\left(\sum_{q=1}^{r} L_q\right)^d\right]. \tag{4.11}$$

Developing with the help of the multinomial identity and using the joint moment function of $L_1, \ldots, L_r$, as obtained from (2.11), gives the result. The stochastic domination property $\mathcal{D}_n(r) \succeq \mathcal{D}_n(r-1)$ follows from the fact that $(\sum_{q=1}^{r} L_q)^d \geq (\sum_{q=1}^{r-1} L_q)^d$ which is maintained when taking the expectation.

(ii) Concerning the mean value of $\mathcal{D}_n(r)$, recalling $1 - \sum_{q=1}^{r} L_q = \prod_{q=1}^{r} \overline{V}_q$ where $\overline{V}_q$, $q = 1, \ldots, r$ are independent with law $\overline{V}_q \overset{d}{\sim} \text{beta}((n-q)\theta, 1+\theta)$ we have from (4.11)

$$\begin{aligned}\boldsymbol{E}_\theta(\mathcal{D}_n(r)) &= \sum_{d \geq 0} \boldsymbol{P}_\theta(\mathcal{D}_{n,k}(r) > d) = \boldsymbol{E}_\theta\left(\frac{1}{1 - \sum_{q=1}^{r} L_q}\right) \\ &= \boldsymbol{E}\left(\frac{1}{\prod_{q=1}^{r} \overline{V}_q}\right) = \prod_{q=1}^{r} \boldsymbol{E}\left(\frac{1}{\overline{V}_q}\right).\end{aligned}$$

The expected value of $\boldsymbol{E}_\theta(\frac{1}{\overline{V}_q})$ is finite and equal to $\frac{(n-q+1)\theta}{(n-q)\theta-1}$ if and only if $(n-q)\theta > 1$. We note that when $r \geq n - \theta^{-1}$ is such that $(n-r)\theta \leq 1$, the expected time separating the visit from the r-th to the $(r+1)$-th new fragment becomes infinitely large. In particular, for values of θ such that $\theta \leq 1/(n-1)$, $\boldsymbol{E}_\theta(\mathcal{D}_n(r)) = \infty$ for each $1 \leq r < n-1$. On the contrary, if $\theta \geq 1$, $\boldsymbol{E}_\theta(\mathcal{D}_n(r)) < \infty$ for each $1 \leq r < n-1$.

Note that, if $(n-r)\theta > 1$, since $\theta > 0$

$$\boldsymbol{E}_\theta(\mathcal{D}_n(r)) = \frac{(n-r+1)\theta}{(n-r)\theta - 1} \boldsymbol{E}_\theta(\mathcal{D}_n(r-1)) > \boldsymbol{E}_\theta(\mathcal{D}_n(r-1)).$$

The expected visiting times sequence to consecutive new species, when these exist (when they are finite), is an increasing one, as conventional wisdom suggests. These results are consistent with the ones of Huillet (2003), obtained in the particular case $\theta = 1$, namely

$$\boldsymbol{E}_{\theta=1}(\mathcal{D}_n(r)) = \prod_{q=1}^{r} \frac{n-q+1}{n-q-1} = \frac{n(n-1)}{(n-r)(n-r-1)},$$

for each $1 \leq r < n-1$. □

Remark. With $\sum_{q=1}^{r} k_q = d$, the following term arising from (4.9)

$$\binom{d}{k_1, \ldots, k_r} \prod_{q=1}^{r} \frac{\Gamma(1+(n-q+1)\theta)\Gamma(1+\theta+k_q)\Gamma((n-q)\theta + k_{q+1} + \cdots + k_r)}{\Gamma(1+\theta)\Gamma((n-q)\theta)\Gamma(1+(n-q+1)\theta + k_q + \cdots + k_r)}$$

is the joint probability $\boldsymbol{P}_\theta(\mathcal{D}_n(r) > d; \mathcal{B}_{n,d}(1) = k_1, \ldots, \mathcal{B}_{n,d}(r) = k_r)$ that $\mathcal{D}_n(r) > d$ and that cell occupancies $\mathcal{B}_{n,d}(q) = k_q$, $q = 1, \ldots, r$ where $\mathcal{B}_{n,d}(q)$ counts the random number of occurrences of the q-th previously visited fragment in a d-sample from $D_n(\theta)$.

The Kingman limit

Consider the situation where $n \uparrow \infty$, $\theta \downarrow 0$ while $n\theta = \gamma > 0$. In this case, we recover part of the results given by Yamato *et al.* (2001) in a broader (two-parameter) model.

As a corollary to Theorem 4.1, we have

COROLLARY 4.1. *Let $p \leq k$. Assume $d_r \geq 1$, $r = 1, \ldots, p-1$, $d_p := k - \sum_{r=1}^{p-1} d_r \geq 1$ and $\overline{d}_r = \sum_{q=1}^{r} d_q$, $r = 1, \ldots, p$, $\overline{d}_0 := 0$. Then, the weak limits $\mathcal{D}_k(r) := \lim_* \mathcal{D}_{n,k}(r)$, $r = 1, \ldots, p-1$ exist and*

(i)

$$\boldsymbol{P}^*_\gamma(\mathcal{D}_k(1) = d_1, \ldots, \mathcal{D}_k(p-1) = d_{p-1}; P_k = p) = \frac{\gamma^p}{(\gamma)_k} \prod_{r=1}^{p} (1 + \overline{d}_{r-1})_{d_r - 1}. \tag{4.12}$$

(ii)

$$\boldsymbol{P}^*_\gamma(\mathcal{D}_k(1) = d_1, \ldots, \mathcal{D}_k(p-1) = d_{p-1} \mid P_k = p) = \frac{1}{s_{k,p}} \prod_{r=1}^{p} (1 + \overline{d}_{r-1})_{d_r - 1}. \tag{4.13}$$

PROOF. Immediate. This is consistent with the result of Yamato *et al.* (2001), Theorem 2, p. 21, putting $\alpha = 0$ in the general two-parameter Pitman class where sampling takes place. This distribution is called *DTGII* by Yamato (1997). □

As a corollary to Theorem 4.2, we get

COROLLARY 4.2. *Let $1 \leq r < k$.*

(i) *With $d_1, \ldots, d_r \geq 1$, $\overline{d}_r < k$,*

$$\boldsymbol{P}^*_\gamma(\mathcal{D}_k(1) = d_1, \ldots, \mathcal{D}_k(r) = d_r) = \frac{\gamma^{r+1}}{(\gamma)_{\overline{d}_r + 1}} \prod_{q=1}^{r} (1 + \overline{d}_{q-1})_{d_q - 1}. \tag{4.14}$$

In particular, if $r = 1$

$$\boldsymbol{P}^*_\gamma(\mathcal{D}_k(1) = d_1) = \gamma \frac{(d_1 - 1)!}{(1+\gamma)_{d_1}}, \quad d_1 \in \{1, \ldots, k-1\} = \frac{(k-1)!}{(1+\gamma)_{k-1}}, \quad d_1 = k, \tag{4.15}$$

showing that $\mathcal{D}_k(1) - 1$ has bounded Waring distribution $bW(k-1, 1+\gamma, 1)$, with

$$\boldsymbol{E}^*_\gamma(\mathcal{D}_k(1)) = 1 + \frac{1}{\gamma - 1} \left(1 - \gamma \frac{k!}{(\gamma)_k}\right). \tag{4.16}$$

(ii) *Let* $1 \le r < k$. *With* $d_1, \ldots, d_r \ge 1$, $\overline{d}_{r-1} =: d \ge r-1$, $\overline{d}_r = d + d_r < k$,

$$\begin{aligned} \boldsymbol{P}^*_\gamma(\mathcal{D}_k(r) = d_r \mid \mathcal{D}_k(1) = d_1, \ldots, \mathcal{D}_k(r-1) = d_{r-1}) \\ = \gamma \frac{(1+d)_{d_r-1}}{(1+\gamma+d)_{d_r}}, \quad \text{if} \quad d_r \in \{1, \ldots, k-d-1\} \\ = \frac{(1+d)_{k-d-1}}{(1+\gamma+d)_{k-d-1}}, \quad \text{if} \quad d_r = k-d, \end{aligned} \tag{4.17}$$

showing that, given $\overline{\mathcal{D}}_k(r-1) := \sum_{q=1}^{r-1} \mathcal{D}_k(q) = d$, $\mathcal{D}_k(r) - 1$ *has bounded Waring distribution* $bW(k-d-1, 1+\gamma+d, 1+d)$. *In particular,*

$$\begin{aligned} \boldsymbol{E}^*_\gamma \left(\mathcal{D}_k(r) \,\middle|\, \sum_{q=1}^{r-1} \mathcal{D}_k(q) = d \right) \\ = 1 + \frac{1+d}{\gamma - 1} \left(1 - \frac{(2+d)_{k-d-1}}{(1+\gamma+d)_{k-d-1}} \right). \end{aligned} \tag{4.18}$$

(iii) *For* $r = 1, \ldots, k-1$, *the cumulative variable* $\overline{\mathcal{D}}_k(r) := \sum_{q=1}^{r} \mathcal{D}_k(q)$ *has distribution*

$$\begin{aligned} \boldsymbol{P}^*_\gamma(\overline{\mathcal{D}}_k(r) = d) = \frac{\gamma^r s_{d,r}}{(1+\gamma)_d}, \quad d = r, r+1, \ldots, k-1 \\ = \frac{\gamma^r s_{d,r}}{(\gamma)_d}, \quad d = k, \end{aligned} \tag{4.19}$$

where $s_{d,r}$ *are the absolute values of first kind Stirling numbers.*

PROOF. Immediate. These are again consistent with results of Yamato *et al.* (2001), as from Theorem 3, Propositions 3 and 4, putting $\alpha = 0$ in their formula. □

Finally, as a corollary to Theorem 4.3, we easily obtain

COROLLARY 4.3. *Let* $r \ge 1$.

(i) *The law of* $\mathcal{D}(r) := \lim_* \mathcal{D}_n(r)$ *is given by*

$$\boldsymbol{P}^*_\gamma(\mathcal{D}(r) > d) = \frac{\gamma^r d!}{(\gamma)_d} \sum_{k_1, \ldots, k_r \ge 0 : \sum_1^r k_q = d} \frac{1}{\prod_{q=1}^r \{\gamma + k_q + \cdots + k_r\}} \tag{4.20}$$

and the following stochastic domination property holds

$$\mathcal{D}(r) \succeq \mathcal{D}(r-1).$$

(ii) *With* $1 \le r$, *the mean value of* $\mathcal{D}(r)$ *reads*

$$\begin{aligned} \boldsymbol{E}^*_\gamma(\mathcal{D}(r)) = \left(\frac{\gamma}{\gamma - 1} \right)^r, \quad \text{if} \quad \gamma > 1 \\ = +\infty, \quad \text{if} \quad 0 < \gamma \le 1. \end{aligned} \tag{4.21}$$

PROOF. Immediate from Theorem 4.3. □

Remark. With $\sum_{q=1}^{r} k_q = d$, $k_q \geq 0$, the following term arising from (4.20)

$$\frac{\gamma^r d!}{(\gamma)_d} \frac{1}{\prod_{q=1}^{r}\{\gamma + k_q + \cdots + k_r\}}$$

is the joint probability $\boldsymbol{P}^*_\gamma(\mathcal{D}(r) > d; \mathcal{B}_d(1) = k_1, \ldots, \mathcal{B}_d(r) = k_r)$ that $\mathcal{D}(r) > d$ and that cell occupancies are $\mathcal{B}_d(q) = k_q$, $q = 1, \ldots, r$, where $\mathcal{B}_d(q)$ counts the random number of occurrences of the q-th previously visited fragment in a d-sample from $GEM(\gamma)$.

REFERENCES

Barrera, J., Huillet, T. and Paroissin, C. (2005). Size-biased permutation of Dirichlet partitions and search-cost distribution, *Probability in the Engineering & Informational Sciences*, **19**(1), 83–97.

Donnelly, P. (1986). Partition structures, Pòlya urns, the Ewens sampling formula and the age of alleles, *Theoretical Population Biology*, **30**, 271–288.

Donnelly, P. (1991). The heaps process, libraries and size-biased permutation, *Journal of Applied Probability*, **28**, 321–335.

Donnelly, P. and Tavaré, S. (1986). The age of alleles and a coalescent, *Advances in Applied Probability*, **18**, 1–19.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles, *Theoretical Population Biology*, **3**, 87–112.

Ewens, W. J. (1990). Population genetics theory—the past and the future, *Mathematical and Statistical Developments of Evolutionary Theory* (ed. S. Lessard), Kluwer, Dordrecht.

Ewens, W. J. (1996). Some remarks on the law of succession, *Athens Conference on Applied Probability and Time Series Analysis (1995), Vol. I*, Lecture Notes in Statistics, **114**, 229–244, Springer, New York.

Huillet, T. (2003). Sampling problems for randomly broken sticks, *Journal of Physics A*, **36**(14), 3947–3960.

Huillet, T. (2005). Sampling formulae arising from random Dirichlet populations, *Communications in Statistics: Theory and Methods* (to appear).

Huillet, T. and Martinez, S. (2003). Sampling from finite random partitions, *Methodology and Computing in Applied Probability*, **5**(4), 467–492.

Kingman, J. F. C. (1975). Random discrete distributions, *Journal of the Royal Statistical Society. Series B*, **37**, 1–22.

Kingman, J. F. C. (1993). *Poisson Processes*, Clarendon Press, Oxford.

Pitman, J. (1996). Random discrete distributions invariant under size-biased permutation, *Advances in Applied Probability*, **28**, 525–539.

Pitman, J. (1999). Coalescents with multiple collisions, *Annals of Probability*, **27**(4), 1870–1902.

Pitman, J. (2002). Poisson-Dirichlet and GEM invariant distributions for split-and-merge transformation of an interval partition, *Combinatorics, Probability and Computing*, **11**(5), 501–514.

Pitman, J. and Yor, M. (1997). The two parameter Poisson-Dirichlet distribution derived from a stable subordinator, *Annals of Probability*, **25**, 855–900.

Sibuya, M. and Yamato, H. (1995). Ordered and unordered random partitions of an integer and the GEM distribution, *Statistics & Probability Letters*, **25**(2), 177–183.

Tavaré, S. and Ewens, W. J. (1997). Multivariate Ewens distribution, *Discrete Multivariate Distributions* (eds. N. L. Johnson, S. Kotz and N. Balakrishnan), **41**, 232–246, Wiley, New York.

Yamato, H. (1997). On the Donnelly-Tavaré-Griffiths formula associated with the coalescent, *Communications in Statistics: Theory and Methods*, **26**(3), 589–599.

Yamato, H., Sibuya, M. and Nomachi, T. (2001). Ordered sample from two-parameter GEM distribution, *Statistics & Probability Letters*, **55**(1), 19–27.