# Sampling formulae arising from random Dirichlet populations

Thierry Huillet

# Sampling formulae arising from random Dirichlet populations

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

October 12, 2004

**Abstract**

Consider the random Dirichlet partition of the interval into $n$ fragments at temperature $\theta > 0$. Some statistical features of this random discrete distribution are recalled, together with explicit results on the law of its size-biased permutation. Using these, pre-asymptotic versions of the Ewens and Donnelly-Tavaré-Griffiths sampling formulae from finite Dirichlet partitions are computed exactly. From these, new proofs of the usual sampling formulae from random proportions with GEM($\gamma$) distribution are supplied, when considering the Kingman limit $n \uparrow \infty$, $\theta \downarrow 0$

while $n\theta = \gamma > 0$ .

Running title: Sampling fom Dirichlet partitions.

Keywords**:** Random discrete distribution, Dirichlet partition, size-biased permutation, GEM, Ewens and Donnelly-Tavaré-Griffiths sampling formulae.

AMS 1991: Primary 60G57, 62E17, Secondary: 60K99, 62E15, 62E20

# 1 Introduction and outline of main results

The joint distribution of unordered (or ordered) frequencies of a sample from random proportions with $GEM\,(\gamma)$ distribution is the Donnelly-Tavaré-Griffiths formula (or the Ewens sampling formulae).

We consider 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. We then recover the Ewens and Donnelly-Tavaré-Griffiths sampling formulae when passing to the Kingman limit $n \uparrow \infty$, $\theta \downarrow 0$, $n\theta = \gamma > 0$, thereby giving new proofs of these famous formulae. For an overview of related problems and further connections between sampling problems, size-biased permutations, combinatorics, random partitions and ran-

dom mappings, we refer to the proceedings of the International Conference [1].

The organization of this manuscript is thus the following. Basic facts on the random Dirichlet partition of the interval into $n$ fragments at temperature $\theta > 0$ are first recalled in Section 2.

In Section 3, explicit results on the law of its size-biased permutation are recalled. These will prove essential when considering the Donnelly-Tavaré-Griffiths sampling formula from Dirichlet partitions in subsequent Subsection 4.3, as given in Theorems 12 and 13.

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 develop are the following: in Proposition 1, it is recalled that the length of an interval containing a random sample is stochastically larger than the typical fragments size from a Dirichlet distribution. Its law is computed. In Theorem 2, the law of the length of the $k-$th fragment in the size-biased permutation is supplied. It is also shown there that the consecutive fragments in the size-biased permutation are arranged in stochastic descending order. In Theorem 3, we give the joint law of the size-biased permutation fragments sizes explicitly (or rather its joint moment function).

The main body of our sampling results is in Section 4. Section 4.1 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.

Section 4.2 concerns the second Ewens sampling formula under the same hypothesis (as a problem of random partitioning of the integers) and Section 4.3 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. Our main results are displayed in Theorems 6, 9 and Theorems 12 and 13 for each of the problems alluded to. Several examples and related facts are supplied.

As corollaries to these Theorems, we show how 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.

## 2 The Dirichlet distribution $D_n(\theta)$

We shall consider the following random partition into $n$ fragments of the unit interval: let $\theta > 0$ be some parameter which we shall interpret as temperature or disorder of the partition. Assume that the random fragments' sizes $\mathbf{S}_n := (S_1, .., S_n)$ (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

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

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

$$\mathbf{E}\left[\prod_{m=1}^{n} S_m^{q_m}\right] = \frac{\Gamma(n\theta)}{\Gamma\left(n\theta + \sum_{m=1}^{n} q_m\right)} \prod_{m=1}^{n} \frac{\Gamma(\theta + q_m)}{\Gamma(\theta)}. \tag{2}$$

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

If this is so, $S_m \overset{d}{=} S_n$, $m = 1, .., n$, independently of $m$ and the individual fragments sizes are all identically distributed (id). Their common density on the interval $(0, 1)$ is given by

$$f_{S_n}(s) = \frac{\Gamma(n\theta)}{\Gamma(\theta)\Gamma((n-1)\theta)} s^{\theta-1}(1-s)^{(n-1)\theta-1}, \tag{3}$$

which is a beta$(\theta, (n-1)\theta)$ density, with mean value $\mathbf{E}(S_n) = 1/n$, variance $\boldsymbol{\sigma}^2(S_n) = \frac{n-1}{n^2(n\theta+1)}$ and moment function $\mathbf{E}[S_n^q] = \frac{\Gamma(\theta+q)\Gamma(n\theta)}{\Gamma(\theta)\Gamma(n\theta+q)}$.

Remark: We recall that a random variable, say $B_{a,b}$, with $B_{a,b} \overset{d}{\sim}$ beta$(a, b)$, has density function $f_{B_{a,b}}(x) := \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$, $a, b > 0$, $x \in [0, 1]$ and moment function $\mathbf{E}\left[B_{a,b}^q\right] = \frac{\Gamma(a+q)\Gamma(a+b)}{\Gamma(a)\Gamma(a+b+q)}$, with $\Gamma(a)$ the Euler's Gamma function. Also, a random variable $T > 0$ with gamma$(\theta)$ distribution has density $f_T(t) := \frac{1}{\Gamma(\theta)} t^{\theta-1} e^{-t}$, $\theta, t > 0$ and moment function $\mathbf{E}[T^q] = \Gamma(\theta + q)/\Gamma(\theta)$.

We also recall that when $\theta = 1$, the partition model Eqs.(1, 2) corresponds to the standard uniform partition model of the interval. □

From Eq. (3), as $n \uparrow \infty$, we next have

$$nS_n \stackrel{d}{\to} \Gamma_{\theta,\theta} \stackrel{d}{\sim} \text{gamma}\,(\theta,\theta)\,, \text{ with density } f_{\Gamma_{\theta,\theta}}(t) = \frac{\theta^\theta}{\Gamma(\theta)} t^{\theta-1} e^{-\theta t},\ t > 0, \tag{4}$$

showing that the sizes of fragments are asymptotically all of order $1/n$.

Consider next the sequence $\mathbf{S}_{(n)} := \left(S_{(m)}; m = 1, .., n\right)$ obtained while ranking the fragments sizes $\mathbf{S}_n$ according to descending sizes, hence with $S_{(1)} > .. > S_{(m)} > .. > S_{(n)}$. The $S_{(m)}$s distribution can hardly be derived in closed form. However, one could prove that, as $n \uparrow \infty$

$$n^{(1+\theta)/\theta} S_{(n)} \stackrel{d}{\to} W_\theta \text{ and } n\theta \left( S_{(1)} - \frac{1}{n\theta} \log\left( n \left(\log n\right)^{\theta-1} \right) \right) \stackrel{d}{\to} G_\theta \tag{5}$$

where $W_\theta$ is a Weibull random variable, $G_\theta$ a Gumbel random variable such that $\mathbf{P}\left(W_\theta > t\right) = \exp\left[-t^\theta / s_\theta\right]$, $t > 0$ and $\mathbf{P}\left(G_\theta \leq t\right) = \exp\left[-s_\theta^{-1} \exp\left(-t\right)\right]$, $t \in \mathbb{R}$, $s_\theta := \Gamma\left(1+\theta\right)\theta^{-\theta} > 0$ a scale parameter.

In the random division of the interval as in Eq. (1) at disorder $\theta$, 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 $\frac{1}{n\theta} \log\left( n \left(\log n\right)^{\theta-1} \right)$. The smaller $\theta$ is, the larger (smaller) the largest (smallest) fragments' size is: hence, the smaller disorder $\theta$ is, the more the values of the $S_m$s are, with high probability, disparate: at low disorder, the size of the largest fragment $S_{(1)}$ tends to dominate the other ones and the range $S_{(1)} - S_{(n)}$ increases when $\theta$ decreases.

On the opposite, large values of $\theta$ correspond to situations in which the range

of fragments' sizes is lower: the fragments' sizes look more homogeneous and distribution Eq. (1) concentrates on its centre. At high disorder, the diversity of the partition is large.

In some applications (see [2] and [3] in the context of the heaps process), $S_m$, $m = 1, .., n$, interpret as the random popularities of a collection of $n$ books arranged on a shelf. If instead of a collection of books, a population of animals from $n$ different species were considered, popularities verbatim interpret as species abundance; see [4] and [5] for such interpretations.

Although $\mathbf{S}_n$ 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 [2], since many interesting statistical features emerge.

# 3 Sampling without replacement and 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 [6]; they were used there to address the following sampling problems from finitely broken sticks (1) What is the sample size if sampling is carried out until the first visit to the smallest fragment? (2) In what order are new fragments being discovered and what is the random number of samples separating the discovery of consecutive

new fragments until exhaustion of the list? They were generalized to all $\theta > 0$ in [7], 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 things self-understandable. They will prove useful in the sequel to derive the Donnelly-Tavaré-Griffiths sampling formula from Dirichlet partitions.

Assume some observer is sampling the unit interval as follows: drop at random points onto this randomly broken interval and record the corresponding numbers of visited fragments. Consider the problem of determining the order in which the various fragments will be discovered in such a sampling process. To avoid revisiting many times the same fragment 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, an estimation of its size is needed. We first do that for the first visited fragment. 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. Its size can itself be estimated 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.

Let $\mathbf{S}_n := (S_1, .., S_n)$ be the random partition of the interval $[0,1]$ considered here with $S_m \overset{d}{=} S_n \overset{d}{\sim} \text{beta}(\theta, (n-1)\,\theta)$, $m = 1, .., n$, $\sum_m S_m = 1$.

Let $U$ be a uniformly distributed random throw on $[0,1]$ and $\mathfrak{L}_n := \mathfrak{L}_n(U)$ the length of the interval of the random partition containing $U$. The distribution of $\mathfrak{L}_n$ is characterized by the conditional probability

$$\mathbf{P}(\mathfrak{L}_n = S_m \mid \mathbf{S}_n) = S_m. \tag{6}$$

In this size-biased picking procedure, long intervals are favored and one expects that $\mathfrak{L}_n \succeq S_n$ in the usual stochastic sense that $\overline{F}_{\mathfrak{L}_n}(s) \geq \overline{F}_{S_n}(s), \forall s \in [0,1]$.

Let us first check that the size of the interval containing $U$ is stochastically larger than the typical fragment's length of the original partition.

## 3.1 The length of a size-biased randomly chosen fragment

From the size-biased picking construction, it follows (see [8], for example) that for all non-negative measurable function $\varphi$ on $[0,1]$,

$$\mathbb{E}_\theta[\varphi(\mathfrak{L}_n)/\mathfrak{L}_n] = \mathbf{E}[\mathbf{E}[\varphi(\mathfrak{L}_n)/\mathfrak{L}_n \mid \mathbf{S}_n]] = \tag{7}$$

$$\sum_{m=1}^{n} \mathbf{E}[\varphi(S_m)/S_m \mathbf{P}(\mathfrak{L}_n = S_m \mid \mathbf{S}_n)] = \sum_{m=1}^{n} \mathbf{E}[\varphi(S_m)].$$

Taking in particular $\varphi(x) = x\mathbf{I}(x > s)$ in Eq. (7), we get the structural distribution

$$\mathbb{P}_\theta[\mathfrak{L}_n > s] := \overline{F}_{\mathfrak{L}_n}(s) = \sum_{m=1}^{n} \mathbf{E}[S_m \mathbf{I}(S_m > s)].$$

Recalling that $S_m \overset{d}{=} S_n$, $m = 1, .., n$, it simplifies to

$$\overline{F}_{\mathfrak{L}_n}(s) = \sum_{m=1}^{n} \int_s^1 t dF_{S_m}(t) = n \int_s^1 t F_{S_n}(dt). \tag{8}$$

**Proposition 1** $\mathfrak{L}_n \overset{d}{\sim}$ *beta*$(1+\theta, (n-1)\theta)$ *and it holds that*

$$\mathfrak{L}_n \succeq S_n. \tag{9}$$

**Proof:** As $S_n \overset{d}{\sim}$ beta$(\theta, (n-1)\theta)$, one can check directly from Eq. (8) that $\mathfrak{L}_n \overset{d}{\sim}$ beta$(1+\theta, (n-1)\theta)$, with $\mathbf{E}(\mathfrak{L}_n) = (1+\theta)/(n\theta+1)$. The likelihood ratio between the two distributions being monotone, the stochastic domination property follows. □

## 3.2 Size-biased permutation of the fragments: one dimensional distribution

Consider the random partition $\mathbf{S}_n$. Let $L_1 := \mathfrak{L}_n$ be the length of the first randomly chosen fragment $M_1$, so with $L_1 := S_{M_1}$ and $\mathbf{P}(M_1 = m_1 \mid \mathbf{S}_n) = S_{m_1}$. 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. We study here this process in some detail.

In the first step of this size-biased picking procedure,

$$\mathbf{S}_n =: \mathbf{S}_n^{(0)} \to (L_1, S_1, .., S_{M_1-1}, S_{M_1+1}, ..S_n)$$

which may be written as $\mathbf{S}_n \to \left(L_1, (1-L_1)\,\mathbf{S}_{n-1}^{(1)}\right)$, with

$$\mathbf{S}_{n-1}^{(1)} := \left(S_1^{(1)}, .., S_{M_1-1}^{(1)}, S_{M_1+1}^{(1)}, ..S_n^{(1)}\right)$$

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 $\mathbf{S}_n$ is the same as that of $(1-L_1)\,\mathbf{S}_{n-1}^{(1)}$ where the $(n-1)-$vector $\mathbf{S}_{n-1}^{(1)} \overset{d}{\sim} D_{n-1}\,(\theta)$ has the distribution of a Dirichlet random partition into $n-1$ fragments (see [9], Chapter 9). Pick next at random an interval in $\mathbf{S}_{n-1}^{(1)}$ 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 $\mathbf{S}_n$ is $\mathbf{L}_n := (L_1, .., L_n)$. We shall put $\mathbf{L}_n = \text{SBP}(\mathbf{S}_n)$.

From this construction, if $(V_1, .., V_{n-1})$ is an independent sample with distribution $V_k \overset{d}{\sim} \text{beta}(1+\theta, (n-k)\,\theta)$, $k = 1, .., n-1$, then,

$$L_k = \prod_{i=1}^{k-1} (1-V_i)\,V_k \ , \ k = 1, .., n-1 \tag{10}$$

$$L_n = 1 - \sum_{k=1}^{n-1} L_k = \prod_{k=1}^{n-1} (1-V_i) \tag{11}$$

is the stick-breaking scheme representation of the size-biased permutation of $\mathbf{S}_n$. 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. From these well-known construction and properties (see [9], Chapter 9, 9.6, [10] and [11], we obtain that the $L_k$s, $k = 1, .., n$ are arranged in stochastically decreasing order. More precisely

**Theorem 2** *(i) The law of $L_k$, for $k = 1, .., n$, is characterized by*

$$\mathbf{E}\left[L_k^q\right] = \prod_{i=1}^{k-1} \mathbf{E}\left[\overline{V}_i^q\right] \mathbf{E}\left[V_k^q\right] =$$

$$\prod_{i=1}^{k-1} \frac{\Gamma\left((n-i)\,\theta + q\right)\Gamma\left((n-i+1)\,\theta + 1\right)}{\Gamma\left((n-i)\,\theta\right)\Gamma\left((n-i+1)\,\theta + 1 + q\right)} \times \frac{\Gamma\left(1+\theta+q\right)\Gamma\left(1+(n-k+1)\,\theta\right)}{\Gamma\left(1+\theta\right)\Gamma\left(1+(n-k+1)\,\theta+q\right)}. \tag{12}$$

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

$$L_k \overset{d}{=} B_{(n-k+1)\theta,1} \cdot L_{k-1},\ k = 2, .., n, \tag{13}$$

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

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

**Proof**: $(i)$ is a direct consequence of the construction, since $\overline{V}_i := 1 - V_i \overset{d}{\sim}$ beta$((n-i)\,\theta, 1+\theta)$, $i = 1, .., k-1$, $V_k \overset{d}{\sim}$ beta$(1+\theta, (n-k)\,\theta)$ are mutually independent. Recalling the expression of the moment function for beta distributions, the corresponding expression of $\mathbf{E}\left[L_k^q\right]$ follows.

$(iii)$ being clearly a consequence of $(ii)$, it remains to prove $(ii)$.

Regrouping terms directly from Eq. (12), we have $\mathbf{E}\left[L_k^q\right] = \mathbf{E}\left[L_{k-1}^q\right] \mathbf{E}\left[B_k^q\right]$ with

$$\mathbf{E}\left[B_k^q\right] = \frac{\Gamma\left((n-k+1)\,\theta + q\right)}{\Gamma\left((n-k+1)\,\theta\right)} \frac{\Gamma\left(1+(n-k+1)\,\theta\right)}{\Gamma\left(1+(n-k+1)\,\theta+q\right)}.$$

This is the moment function of a beta$((n-k+1)\,\theta, 1)$ distributed random variable. □

Let us now compute the joint distribution of the size-biased permutation $\mathbf{L}_n$ of $\mathbf{S}_n$. We shall say in the sequel that, if $\mathbf{L}_n = \text{SBP}(\mathbf{S}_n)$, then $\mathbf{L}_n \overset{d}{\sim} SBD_n(\theta)$ assuming that $\mathbf{S}_n \overset{d}{\sim} D_n(\theta)$.

## 3.3 Joint law of the size-biased permutation of a Dirichlet partition

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

$$\mathbf{P}\left(M_1^{'} = m_1, .., M_k^{'} = m_k \mid \mathbf{S}_n\right) = \prod_{i=1}^{k-1} \frac{S_{m_i}}{1 - \sum_{l=1}^{i} S_{m_l}} S_{m_k}. \tag{14}$$

Let us now compute the joint distribution of the size-biased permutation $\mathbf{L}_n$ of $\mathbf{S}_n$ with $\mathbf{L}_n \overset{d}{\sim} SBD_n(\theta)$ and $\mathbf{S}_n \overset{d}{\sim} D_n(\theta)$. First, we have

$$(L_1, .., L_n) = \left(S_{M_1^{'}}, .., S_{M_n^{'}}\right), \tag{15}$$

and consequently

$$\mathbf{P}\left(L_1 = S_{m_1}, .., L_n = S_{m_n} \mid \mathbf{S}_n\right) = \prod_{k=1}^{n-1} \frac{S_{m_k}}{1 - \sum_{l=1}^{k} S_{m_l}} S_{m_n}. \tag{16}$$

We shall now rather consider the joint moment function of the random size-biased permutation $\mathbf{L}_n = (L_1, .., L_n)$. We can prove the following result

**Theorem 3** *The joint moment function of the SBP* $\mathbf{L}_n = (L_1, .., L_n) \overset{d}{\sim} SBD_n(\theta)$ *reads*

$$\mathbf{E}\left[\prod_{k=1}^{n} L_k^{q_k}\right] = \sum_{\{m_1 \neq .. \neq m_n\}} \mathbf{E}\left[\prod_{k=1}^{n-1} \frac{S_{m_k}^{q_k+1}}{1-\sum_{l=1}^{k} S_{m_l}} S_{m_n}^{q_n+1}\right] = \tag{17}$$

$$\prod_{k=1}^{n-1}\left\{\frac{\Gamma(1+(n-k+1)\theta)}{\Gamma(1+\theta)\Gamma((n-k)\theta)}\frac{\Gamma(1+\theta+q_k)\Gamma((n-k)\theta+q_{k+1}+..+q_n)}{\Gamma(1+(n-k+1)\theta+q_k+..+q_n)}\right\}.$$

**Proof**: The quantity $\mathbf{E}\left[\prod_{k=1}^{n} L_k^{q_k}\right]$ has the expression given by the first equality as a result of (16), summing over all permutations $\{m_1, ..., m_n\}$ of $\{1, ..., n\}$ and taking the average over $\mathbf{S}_n$.

Next, we observe from Eqs. (10, 11) and the independence of the $V_k$s that

$$\mathbf{E}\left[\prod_{k=1}^{n} L_k^{q_k}\right] = \mathbf{E}\left[\prod_{k=1}^{n}\prod_{i=1}^{k-1} \overline{V}_i^{q_k} V_k^{q_k}\right] = \prod_{k=1}^{n-1} \mathbf{E}\left[V_k^{q_k}\overline{V}_k^{q_{k+1}+..+q_n}\right], \tag{18}$$

with $V_k \overset{d}{\sim} \text{beta}(1+\theta, (n-k)\theta)$, $\overline{V}_k \overset{d}{\sim} \text{beta}((n-k)\theta, 1+\theta)$, $k = 1, .., n-1$.

Finally, suppose $V \overset{d}{\sim} \text{beta}(a, b)$. Then, with $\overline{V} := 1 - V$, it holds that

$$\mathbf{E}\left[V^{q_1}\overline{V}^{q_2}\right] = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^1 v^{a+q_1-1}(1-v)^{b+q_2-1}\,dv$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a+q_1)\Gamma(b+q_2)}{\Gamma(a+b+q_1+q_2)}.$$

Adapting this computation, recalling that $V_k \overset{d}{\sim} \text{beta}(1+\theta, (n-k)\theta)$, the quantity $\mathbf{E}\left[V_k^{q_k}\overline{V}_k^{q_{k+1}+..+q_{n-1}}\right]$ appearing in Eq. (18), has the expression displayed inside the product in the second part of (17). □

Remark: We shall borrow the physical image to the heaps process (see [2] and [12]). Books' popularities are assumed to satisfy $\mathbf{S}_n \overset{d}{\sim} D_n(\theta)$. When a

book is demanded, it is removed and replaced (before a next demand) to the top of the shelf, other books being shifted accordingly; successive demands are independent. Iterating this heaps process (as a recurrent positive Markov chain over the set of permutations), there is intuitively a tendency, when the system has reached equilibrium, to find more popular books to the top of the heap. At equilibrium indeed (see [3] and references therein to Dies, Hendricks and Letac' works), books' popularities are given by $\mathbf{L}_n = \mathrm{SBP}(\mathbf{S}_n) \overset{d}{\sim} SBD_n(\theta)$ and result $(iii)$ in Theorem 2 stating that $L_1 \succeq .. \succeq L_n$ confirms and gives statistical sense to this intuition. Note from this that $\mathbf{L}_n = \mathrm{SBP}(\mathbf{L}_n)$ ($\mathbf{L}_n$ is invariant under size-biased permutation) and that $\mathbf{L}_n = \mathrm{SBP}\big(\mathbf{S}_{(n)}\big)$ since $\mathbf{S}_{(n)}$ is simply obtained from $\mathbf{S}_n$ while rearranging its components in descending order, observing that the sampling process is blind to the mutual fragments' positions, being only sensitive to their sizes. For an application of these results to the search-cost distribution in a heap under a move-to-front rule, see [7].

## The Kingman limit

Consider the situation where $n \uparrow \infty$, $\theta \downarrow 0$ while $n\theta = \gamma > 0$. Such an asymptotics was first considered by [2]; we shall "star" the results (as in $\underset{*}{\overset{d}{\rightarrow}}$) when referring to such an asymptotics. As noted by this author, $\mathbf{S}_n \overset{d}{\sim} D_n(\theta)$ itself has no non-degenerate limit.

When $k = o(n)$, recalling $V_k \overset{d}{\sim} \mathrm{beta}(1+\theta, (n-k)\theta)$, we have $V_k \underset{*}{\overset{d}{\rightarrow}} V_k^* \overset{d}{\sim} \mathrm{beta}(1, \gamma)$ and the $SBD_n(\theta)$ distribution converges weakly from Eqs. (10, 11) to

a Griffiths-Engen-McCloskey or $GEM(\gamma)$ distribution. Namely, $(L_1, .., L_n) \xrightarrow[*]{d}$ $(L_1^*, .., L_k^*, ..) =: \mathbf{L}^*$ where

$$L_k^* = \prod_{i=1}^{k-1} \overline{V}_i^* V_k^*, \; k \geq 1. \tag{19}$$

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

The model (19) 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 [13] and the references therein for example; this model and related ones are also fundamental in Probability Theory; see [14], [15], [16] and [17].

# 4 Dirichlet partitions: sampling formulae for unordered and ordered sequences

Ewens' sampling formula (ESF) gives the distribution of alleles (different types of genes) in a $k-$sample from the Poisson-Dirichlet process $PD(\gamma)$. Alter-

natively, 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, $i = 0, .., k$. When doing this while assuming the species have random frequencies following $GEM\,(\gamma)$ distribution, we are led to a second Ewens Sampling Formula.

Our goal here is first to supply exact expressions of both first and second Ewens sampling formulae, when sampling is from finite Dirichlet random partitions, assuming sampled population to be made of $n$ elements. Similarly, we shall supply a DTG formula, when sampling is from finite Dirichlet random partitions. We shall then show in each case that these sampling formulae give both ESF and DTG formulae when passing to the Kingman limit, thereby giving a new proof of these well-known results under the $GEM$ model (see [18] for further results). To derive the pre-asymptotic DTG sampling formula, the joint law of the size-biased permutation of a Dirichlet partition will be needed.

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

$$\mathbf{P}\left(M = m \mid S_1, .., S_n\right) = S_m,\ m \in \{1, .., n\} \tag{20}$$

and

$$\mathbb{P}_\theta\left(M = m\right) := \mathbf{E}\left[\mathbf{P}\left(M = m \mid \mathbf{S}_n\right)\right] = \mathbf{E}S_m = \frac{1}{n}. \tag{21}$$

Let $\mathcal{B}_{n,k}\left(m\right) = \sum_{l=1}^{k} \mathbf{I}\left(M_l = m\right)$ count the random number of occurrences of fragment $m$ in the $k-$sample. With $\sum_{m=1}^{n} \mathcal{B}_{n,k}\left(m\right) = k$, $\mathcal{B}_{n,k}\left(m\right)$ has Binomial distribution and, as a result, for any $p \in \{1, ..n\}$ and any sequence $1 \leq n_1 < .. < n_p \leq n$, the following multinomial distribution representation holds:

**Lemma 4** *With $(k_1, .., k_p) \in \{0, .., k\}^p$, such that $\sum_{q=1}^{p} k_q \leq k$, it holds*

$$\begin{aligned}
&\mathbf{P}\left(M_1, .., M_k \in \{n_1, .., n_p\}; \mathcal{B}_{n,k}\left(n_1\right) = k_1, .., \mathcal{B}_{n,k}\left(n_p\right) = k_p \mid \mathbf{S}_n\right) \\
&= \frac{k!}{\prod_{q=1}^{p} k_q! \left(k - \sum_{q=1}^{p} k_q\right)!} \prod_{q=1}^{p} S_{n_q}^{k_q} \cdot \left(1 - \sum_{q=1}^{p} S_{n_q}\right)^{k - \sum_{q=1}^{p} k_q}.
\end{aligned} \tag{22}$$

Let $P_{n,k} := \sum_{m=1}^{n} \mathbf{I}\left(\mathcal{B}_{n,k}\left(m\right) > 0\right)$ count the number of distinct fragments which have been visited in the $k-$sampling process. If now $(k_1, .., k_p) \in \{1, .., k\}^p$ are such that $\sum_{q=1}^{p} k_q = k$, it follows from the above Lemma 4 that

**Corollary 5**

$$\mathbf{P}\left(M_1, .., M_k \in \{n_1, .., n_p\}; \mathcal{B}_{n,k}\left(n_1\right) = k_1, .., \mathcal{B}_{n,k}\left(n_p\right) = k_p; P_{n,k} = p \mid \mathbf{S}_n\right)$$

$$= \frac{k!}{\prod_{q=1}^{p} k_q!} \prod_{q=1}^{p} S_{n_q}^{k_q}, \tag{23}$$

*where $p$ satisfies $1 \leq p \leq n \wedge k$.*

Averaging over $\mathbf{S}_n$, observing that the sample function $\mathbf{S}_n \to \prod_{q=1}^{p} S_{n_q}^{k_q}$ is homogeneous of degree $k$, applying $(ii)$ of Theorem 1 page 471 of [19], we have

$$\mathbf{E}\left[\prod_{q=1}^{p} S_{n_q}^{k_q}\right] = \frac{\Gamma\left(n\theta\right)}{\Gamma\left(n\theta + k\right)} \mathbf{E}\left[\prod_{q=1}^{p} T_q^{k_q}\right] = \frac{\Gamma\left(n\theta\right)}{\Gamma\left(n\theta + k\right)} \prod_{q=1}^{p} \mathbf{E}\left[T_q^{k_q}\right]$$

where $(T_q;\ q = 1, .., p)$ are iid random variables on $(0, \infty)$ with $T_q \overset{d}{\sim} \text{gamma}(\theta)$, $q = 1, .., p$ and $\mathbf{E}\left[T_q^{k_q}\right] = \frac{\Gamma(\theta + k_q)}{\Gamma(\theta)} =: \left(\theta\right)_{k_q}$. Therefore, we obtain

$$\begin{aligned} & \mathbb{P}_\theta\left(M_1, .., M_k \in \{n_1, .., n_p\}; \mathcal{B}_{n,k}\left(n_1\right) = k_1, .., \mathcal{B}_{n,k}\left(n_p\right) = k_p; P_{n,k} = p\right) \\ = & \frac{k!}{\prod_{q=1}^{p} k_q!} \frac{\Gamma\left(n\theta\right)}{\Gamma\left(n\theta + k\right)} \prod_{q=1}^{p} \left(\theta\right)_{k_q}. \end{aligned}$$

The above probability is independent of the sequence $n_1 < .. < n_p$. As there are $\binom{n}{p}$ such sequences, if $\mathcal{B}_{n,k}\left(q\right)$, $q = 1, .., p$, stand for the numbers of animals of species $q$ where the $P_{n,k}$ species observed were labelled in an arbitrary way (independently of the sampling mechanism), we finally obtain

**Theorem 6** *(i) It holds*

$$\mathbb{P}_\theta\left(\mathcal{B}_{n,k}(1)=k_1,..,\mathcal{B}_{n,k}(p)=k_p;P_{n,k}=p\right)$$
$$=\binom{n}{p}\frac{k!}{\prod_{q=1}^{p}k_q!}\frac{1}{(n\theta)_k}\prod_{q=1}^{p}(\theta)_{k_q}. \tag{24}$$

*(ii) With*

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

*the Bell polynomials, we have*

$$\mathbb{P}_\theta(P_{n,k}=p)=\frac{n!}{(n-p)!}\frac{1}{(n\theta)_k}B_{k,p}\left((\theta)_1,(\theta)_2...\right). \tag{25}$$

*(iii) It holds that*

$$\mathbb{P}_\theta\left(\mathcal{B}_{n,k}(1)=k_1,..,\mathcal{B}_{n,k}(p)=k_p\mid P_{n,k}=p\right)$$
$$=\frac{k!}{p!}\frac{1}{B_{k,p}\left((\theta)_1,(\theta)_2...\right)}\prod_{q=1}^{p}\frac{(\theta)_{k_q}}{k_q!}. \tag{26}$$

**Proof**: Part $(i)$ has already been proven. Part $(ii)$ is not obvious at this stage. It will be proven rigorously as a consequence of the second Ewens Formula for Dirichlet partitions derived in Theorem 9 of Subsection 4.2. Part $(iii)$ is a consequence of $(i)$ and $(ii)$. The problem consisting in calculating the probability $\mathbb{P}_\theta(P_{n,k}=p)$ is known as the Chinese Restaurant Problem. □

Example: As a particular example, we consider the critical case $\theta=1$. In this case, the above formula simplifies to

$$\mathbb{P}_1\left(\mathcal{B}_{n,k}(1)=k_1,..,\mathcal{B}_{n,k}(p)=k_p;P_{n,k}=p\right)=\frac{\binom{n}{p}}{\binom{n+k-1}{k}},$$

which is independent of the cell occupancies $(k_1, .., k_p)$ (the probability is uniform).

As there are $\binom{k-1}{p-1}$ sequences $k_q \geq 1$, $q = 1, .., p$ satisfying $\sum k_q = k$, we get

$$\mathbb{P}_1 \left( P_{n,k} = p \right) = \frac{\binom{n}{p} \binom{k-1}{p-1}}{\binom{n+k-1}{k}}, \ p = 1, .., n \wedge k.$$

As a result,

$$\mathbb{P}_1 \left( \mathcal{B}_{n,k} \left( 1 \right) = k_1, .., \mathcal{B}_{n,k} \left( p \right) = k_p \mid P_{n,k} = p \right) = \frac{1}{\binom{k-1}{p-1}}. \ \square$$

Remark (the law of succession): To be complete, we would like to briefly revisit a related question raised in [11] and [20], concerning the law of succession.

$(i)$ Consider Eq (23) and let

$$\mathbf{P} \left( M_1 ... M_k \in \{ n_1, .., n_p \} ; M_{k+1} \text{ new}; \mathcal{B}_{n,k} \left( n_1 \right) = k_1, .. \mathcal{B}_{n,k} \left( n_p \right) = k_p; P_{n,k} = p \mid \mathbf{S}_n \right)$$

be the probability that a $(k+1)^{th}$ sample is not amongst the ones $\{ n_1, .., n_p \}$ previously encountered (and so is new), given $\mathbf{S}_n$. From Eq (23), this probability may be written as

$$\frac{k!}{\prod_{q=1}^{p} k_q!} \prod_{q=1}^{p} S_{n_q}^{k_q} \left( 1 - \sum_{q=1}^{p} S_{n_q} \right).$$

The function $\mathbf{S}_n \to \prod_{q=1}^{p} S_{n_q}^{k_q} \left( 1 - \sum_{q=1}^{p} S_{n_q} \right)$ is homogeneous with degree $k+1$. Taking the average over $\mathbf{S}_n$, applying the usual trick, this probability reads

$$\frac{k!}{\prod_{q=1}^{p} k_q!} \frac{\Gamma \left( n\theta \right)}{\Gamma \left( n\theta + k + 1 \right)} \prod_{q=1}^{p} \left( \theta \right)_{k_q} \times \left( n - p \right) \theta.$$

Summing over the sequences $\{n_1, .., n_p\}$ and conditioning, Eq (24) yields

$$\mathbf{P}_\theta \left( M_{k+1} \text{ is new} \mid \mathcal{B}_{n,k}(1) = k_1, .., \mathcal{B}_{n,k}(p) = k_p; P_{n,k} = p \right) = \frac{(n-p)\,\theta}{n\theta + k}, \quad (27)$$

which is independent of cell occupancies $k_1, .., k_p$ but depends on the number $p$ of distinct fragments already visited by the $k-$sample. Note that $p \leq (n-1) \wedge k$ if this probability is to be strictly positive.

$(ii)$ Similarly, consider Eq (23) and, with $n_r \in \{n_1, .., n_p\}$, let

$$\mathbf{P}\left( M_1 ... M_k \in \{n_1, .., n_p\}; M_{k+1} = n_r; \mathcal{B}_{n,k}(n_1) = k_1, .. \mathcal{B}_{n,k}(n_p) = k_p; P_{n,k} = p \mid \mathbf{S}_n \right)$$

be the probability that the $(k+1)^{th}$ sample is one from the previously encountered fragment already visited $k_r$ times, given $\mathbf{S}_n$. This probability is also

$$\frac{k!}{\prod_{q=1}^{p} k_q!} \prod_{\substack{q=1 \\ q \neq r}}^{p} S_{n_q}^{k_q} \times S_{n_r}^{k_r+1}.$$

Averaging over $\mathbf{S}_n$, summing over the sequences $\{n_1, .., n_p\}$ and conditioning, we easily get, proceeding as in $(i)$

$$\mathbf{P}_\theta \left( M_{k+1} \in \text{ species seen } k_r \text{ times} \mid \mathcal{B}_{n,k}(1) = k_1, .., \mathcal{B}_{n,k}(p) = k_p; P_{n,k} = p \right)$$
$$= \frac{\theta + k_r}{n\theta + k}, \quad (28)$$

which is independent of $k_q$, $q \in \{1, .., p\} \setminus \{r\}$, and also of $p$. $\square$

## The Kingman limit

Consider the situation where $n \uparrow \infty$, $\theta \downarrow 0$ while $n\theta = \gamma > 0$. We recover a result first given in [21] in a way which constitutes a new proof of the ESF. Indeed, we have

**Corollary 7** *In the Kingman limit,* $\mathbb{P}_\theta\left(\mathcal{B}_{n,k}\left(1\right)=k_1,..,\mathcal{B}_{n,k}\left(p\right)=k_p;P_{n,k}=p\right)$ *converges to*

$$\mathbb{P}_\gamma^*\left(\mathcal{B}_k\left(1\right)=k_1,..,\mathcal{B}_k\left(p\right)=k_p;P_k=p\right)=\frac{k!}{p!}\frac{\gamma^p}{\left(\gamma\right)_k\prod_{q=1}^p k_q}. \tag{29}$$

**Proof**: From Stirling formula, we have $\binom{n}{p}\sim\frac{n^p}{p!}$ and $\frac{\Gamma(n\theta)}{\Gamma(n\theta+k)}\sim\frac{1}{(\gamma)_k}$, where $\left(\gamma\right)_k:=\gamma\left(\gamma+1\right)..\left(\gamma+k-1\right)$.

Furthermore, $\prod_{q=1}^p\frac{\Gamma(\theta+k_q)}{\Gamma(\theta)}=\theta^p\prod_{q=1}^p\frac{\Gamma(\theta+k_q)}{\Gamma(1+\theta)}\sim\theta^p\prod_{q=1}^p\left(k_q-1\right)!$. □

Summing over $k_1,..,k_p$ satisfying $k_q\geq 1$, $q=1,..,p$ and $\sum k_q=k$ gives the limiting probability $\mathbb{P}_\gamma^*\left(P_k=p\right)$ that there are $p\leq k$ distinct species visited in the $k-$sample. So

**Corollary 8** *With* $s_{k,p}$ *the absolute value of the first kind Stirling numbers, it holds that*

$$\mathbb{P}_\gamma^*\left(P_k=p\right)=\frac{\gamma^p s_{k,p}}{\left(\gamma\right)_k},\ p=1,..,k \tag{30}$$

*and*

$$\mathbb{P}_\gamma^*\left(\mathcal{B}_k\left(1\right)=k_1,..,\mathcal{B}_k\left(p\right)=k_p\mid P_k=p\right)=\frac{k!}{p!}\frac{1}{s_{k,p}\prod_{q=1}^p k_q}. \tag{31}$$

**Proof**: One of the expressions of $s_{k,p}:=\left[\gamma^p\right]\left(\gamma\right)_k$ (as the coefficient of $\gamma^p$ in the development of $\left(\gamma\right)_k$) is

$$s_{k,p}=\frac{k!}{p!}\sum\frac{1}{\prod_{q=1}^p k_q},$$

where the summation runs over the integers $k_q\geq 1$, $q=1,..,p$ satisfying $\sum k_q=k$ (see [21]). □

Remark (the law of succession)**:** In the Kingman limit, the probabilities displayed in Examples (27) and (28) converge respectively to

$$\frac{\gamma}{\gamma+k} \text{ and } \frac{k_r}{\gamma+k}. \tag{32}$$

These arise in the Pŏlya urn model [22]. □

## 4.2 The second Ewens formula for Dirichlet populations

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

$$\mathcal{A}_{n,k}(i) = \#\{m \in \{1,..,n\} : \mathcal{B}_{n,k}(m) = i\} = \sum_{m=1}^{n} \mathbf{I}(\mathcal{B}_{n,k}(m) = i). \tag{33}$$

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. Besides, $\sum_{i=1}^{k} i\mathcal{A}_{n,k}(i) = k$ is the sample size.

The vector $(\mathcal{A}_{n,k}(1),..,\mathcal{A}_{n,k}(k))$ is called the fragments vector count (or the species vector count in biology [5]). In this case, we have

**Theorem 9** *(i) For any $a_i \geq 0$, $i = 1,..,k$ satisfying $\sum_{i=1}^{k} ia_i = k$ and $\sum_{i=1}^{k} a_i = p$, we have*

$$\mathbb{P}_\theta(\mathcal{A}_{n,k}(1) = a_1,..,\mathcal{A}_{n,k}(k) = a_k; P_{n,k} = p) \tag{34}$$

$$= \frac{n!}{(n-p)!} \frac{k!}{\prod_{i=1}^{k} i!^{a_i} a_i!} \frac{\Gamma(n\theta)}{\Gamma(n\theta+k)} \prod_{i=1}^{k} (\theta)_i^{a_i}. \tag{35}$$

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

$$\mathbb{P}_\theta(P_{n,k} = p) = \frac{n!}{(n-p)!}\frac{\Gamma(n\theta)}{\Gamma(n\theta+k)} B_{k,p}((\theta)_1, (\theta)_2 ...). \tag{36}$$

*(iii) It holds*

$$\mathbb{P}_\theta(\mathcal{A}_{n,k}(1) = a_1, .., \mathcal{A}_{n,k}(k) = a_k \mid P_{n,k} = p) \tag{37}$$

$$= \frac{k!}{B_{k,p}((\theta)_1, (\theta)_2 ...)} \prod_{i=1}^{k} \frac{(\theta)_i^{a_i}}{i!^{a_i} a_i!}. \tag{38}$$

**Proof:** Part $(i)$ follows from Proposition 5.1 of [23]; see also Proposition 7 of [19].

$(ii)$ Consider the Bell polynomials (see [24] pages $144 - 147$, Tome 1)

$$B_{k,p}(x_1, x_2, ..) = \sum \frac{k!}{\prod_{i=1}^{k} i!^{a_i} a_i!} \prod_{i=1}^{k} x_i^{a_i},$$

where the summation runs over the integers $a_i \geq 0$, $i = 1, .., k$ satisfying $\sum_{i=1}^{k} i a_i = k$ and $\sum_{i=1}^{k} a_i = p$. It holds that, with monomials $x_i$ particularized to $x_i = (\theta)_i$

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

This result clearly solves the same problem raised in part $(ii)$ of Theorem 6, as both Eqs (24) and (34) share the same marginal $\mathbb{P}_\theta(P_{n,k} = p)$. Part $(iii)$ results from normalization. □

Example: As a particular example, we consider the case $\theta = 1$. In this case, the above formula simplifies to

$$\mathbb{P}_1(\mathcal{A}_{n,k}(1) = a_1, .., \mathcal{A}_{n,k}(k) = a_k; P_{n,k} = p) = \frac{p!\binom{n}{p}}{\binom{n+k-1}{k}} \frac{1}{\prod_{i=1}^{k} a_i!}.$$

Considering Bell polynomials $B_{k,p}(x_1, x_2, ...)$ where monomials $x_i$ are particularized to $x_i = (1)_i = i!$, we have

$$\sum \frac{k!}{\prod_{i=1}^{k} a_i!} = \frac{k!}{p!}\binom{k-1}{p-1}.$$

As a result, we get, as expected

$$\mathbb{P}_1(P_{n,k} = p) = \frac{\binom{n}{p}\binom{k-1}{p-1}}{\binom{n+k-1}{k}},\ p = 1, .., n \wedge k.$$

Furthermore, the conditional distribution reads

$$\mathbb{P}_1(\mathcal{A}_{n,k}(1) = a_1, .., \mathcal{A}_{n,k}(k) = a_k \mid P_{n,k} = p) = \frac{p!}{\binom{k-1}{p-1}} \frac{1}{\prod_{i=1}^{k} a_i!}. \square$$

The Kingman limit

Consider the situation where $n \uparrow \infty$, $\theta \downarrow 0$ while $n\theta = \gamma > 0$. We shall recover the celebrated Ewens Sampling Formula, [21]. Indeed, we have

**Corollary 10** *In the Kingman limit,* $\mathbb{P}_\theta(\mathcal{A}_{n,k}(1) = a_1, .., \mathcal{A}_{n,k}(k) = a_k; P_{n,k} = p)$ *converges to*

$$\mathbb{P}^*_\gamma(\mathcal{A}_k(1) = a_1, .., \mathcal{A}_k(k) = a_k; P_k = p) = \frac{k!\gamma^p}{(\gamma)_k \prod_{i=1}^{k} i^{a_i} a_i!}. \tag{39}$$

**Proof:**

We have $\frac{n!}{(n-p)!} \sim n^p$, $\frac{\Gamma(n\theta)}{\Gamma(n\theta+k)} \sim \frac{1}{(\gamma)_k}$ where $(\gamma)_k := \gamma(\gamma+1)..(\gamma+k-1)$ and $\prod_{i=1}^{k}\left(\frac{\Gamma(\theta+i)}{\Gamma(\theta)}\right)^{a_i} = \theta^p \prod_{i=1}^{k}\left(\frac{\Gamma(\theta+i)}{\Gamma(1+\theta)}\right)^{a_i} \sim \theta^p \prod_{i=1}^{k}(i-1)!$. $\square$

Summing over $a_i \geq 0$, $i = 1, .., k$ satisfying $\sum_{i=1}^{k} i a_i = k$ and $\sum_{i=1}^{k} a_i = p$ gives the limiting probability $\mathbb{P}_{\gamma}^{*}(P_k = p)$ that there are $p \leq k$ distinct species visited in the $k-$sample. We find

**Corollary 11** *With $s_{k,p}$ the absolute value of the first kind Stirling numbers, it holds that*

$$\mathbb{P}_{\gamma}^{*}(P_k = p) = \frac{\gamma^p s_{k,p}}{(\gamma)_k},\ p = 1, .., k \tag{40}$$

*and*

$$\mathbb{P}_{\gamma}^{*}(\mathcal{A}_k(1) = a_1, .., \mathcal{A}_k(p) = a_k \mid P_k = p) = \frac{k!}{s_{k,p} \prod_{i=1}^{k} i^{a_i} a_i!}. \tag{41}$$

**Proof:** Another expression of $s_{k,p} := [\gamma^p](\gamma)_k$ (as the coefficient of $\gamma^p$ in $(\gamma)_k$) is in terms of Bell polynomials $B_{k,p}(x_1, x_2, ..)$ when monomials $x_i$ are particularized to $x_i = (i-1)!$ (see [24], pages $146-147$, Volume 1). This may also be seen directly from Eq. (36), when passing to the Kingman limit, observing that $B_{k,p}((\theta)_1, (\theta)_2, ..) \sim_{\theta \downarrow 0} \theta^p B_{k,p}(0!, 1!, 2!, ..)$. $\square$

## 4.3 Donnelly-Tavaré-Griffiths sampling formulae for Dirichlet partitions

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 the $k-$sample and let $m_1 \neq m_2 \neq ... \neq m_p$ denote the ordered

number of the first, second,..., $p^{th}$ distinct animals sampled from the size-$p$ sub-sample of $\mathbf{S}_n$ corresponding to the $p$ distinct fragments which were visited. Let $\mathfrak{C}_{n,k}(q)$, $q = 1, .., p$ be the number of animals of $q^{th}$ species to appear and $P_{n,k} := \sum_{m=1}^{n} \mathbf{I}(\mathfrak{C}_{n,k}(m) > 0)$ be the total number of distinct visited species. We have

**Theorem 12** *(i) For any $k_q \geq 1$, $q = 1, .., p$ satisfying $\sum k_q = k$ and any $p = 1, .., n \wedge k$, it holds*

$$\mathbb{P}_\theta(\mathfrak{C}_{n,k}(1) = k_1, .., \mathfrak{C}_{n,k}(p) = k_p; P_{n,k} = p) \tag{42}$$

$$= \binom{n}{p} \frac{k!}{\prod_{q=1}^{p} k_q!} \frac{\Gamma(n\theta)\Gamma(p\theta + k)}{\Gamma(n\theta + k)\Gamma(p\theta)} \times \tag{43}$$

$$\prod_{q=1}^{p-1} \left\{ \frac{\Gamma(1 + (p-q+1)\theta)\Gamma(1 + \theta + k_q)}{\Gamma(1+\theta)\Gamma((p-q)\theta)} \frac{\Gamma((p-q)\theta + k_{q+1} + .. + k_p)}{\Gamma(1 + (p-q+1)\theta + k_q + .. + k_p)} \right\}.$$

*(ii) It holds that*

$$\mathbb{P}_\theta(P_{n,k} = p) = \frac{n!}{(n-p)!} \frac{\Gamma(n\theta)}{\Gamma(n\theta + k)} B_{k,p}((\theta)_1, (\theta)_2 ...). \tag{44}$$

*(iii) The conditional distribution given $P_{n,k} = p$ reads*

$$\mathbb{P}_\theta(\mathfrak{C}_{n,k}(1) = k_1, .., \mathfrak{C}_{n,k}(p) = k_p \mid P_{n,k} = p) \tag{45}$$

$$= \frac{k!}{p!} \frac{\Gamma(p\theta + k)}{\Gamma(p\theta) B_{k,p}((\theta)_1, (\theta)_2 ...)} \frac{1}{\prod_{q=1}^{p} k_q!} \times \tag{46}$$

$$\prod_{q=1}^{p-1} \left\{ \frac{\Gamma(1 + (p-q+1)\theta)\Gamma(1 + \theta + k_q)}{\Gamma(1+\theta)\Gamma((p-q)\theta)} \frac{\Gamma((p-q)\theta + k_{q+1} + .. + k_p)}{\Gamma(1 + (p-q+1)\theta + k_q + .. + k_p)} \right\}.$$

**Proof:** *(i)* Let $\{n_1, ...n_p\}$ be any sequence of integers satisfying $1 \leq n_1 < .. < n_p \leq n$. There are $\binom{n}{p}$ such sequences. Given $\mathbf{S}_n$, the probability that the

$k-$sample falls in $\left(S_{n_1},..,S_{n_p}\right)$ is $\left(\sum_1^p S_{n_q}\right)^k$. The function $\mathbf{S}_n \to \left(\sum_1^p S_{n_q}\right)^k$ is homogeneous with degree $k$ and if $T_q \overset{d}{\sim} \text{gamma}(\theta)$, $q = 1,..,p$, are iid, $\mathbf{E}\left[\left(\sum_1^p T_q\right)^k\right] = \frac{\Gamma(p\theta+k)}{\Gamma(p\theta)}$. Averaging over $\mathbf{S}_n$, the probability of this event is thus given by

$$\mathbf{E}\left[\left(\sum_1^p S_{n_q}\right)^k\right] = \frac{\Gamma(n\theta)}{\Gamma(n\theta+k)}\frac{\Gamma(p\theta+k)}{\Gamma(p\theta)}.$$

Relabel $\{n_1,...n_p\}$ as $\{1,..,p\}$ and consider the new partition of the interval $(S_1,..,S_p;\mathbf{S}_{n-p})$, moving $S_1,..,S_p$ to the front and shifting the $n-p$ remaining terms accordingly to form $\mathbf{S}_{n-p}$. Consider then the $p-$partition of the unity $\mathbf{\Sigma}_p$ defined upon scaling by: $(S_1,..,S_p;\mathbf{S}_{n-p}) =: (\sum_1^p S_q \cdot \mathbf{\Sigma}_p;\mathbf{S}_{n-p})$. It holds that $\mathbf{\Sigma}_p$ and $\sum_1^p S_q$ are independent; furthermore $\mathbf{\Sigma}_p \overset{d}{\sim} D_p(\theta)$. This results from the well-known fact that for Dirichlet partitions, $S_m = T_m/\sum_1^n T_k$, $m = 1,..,n$ where $(T_k;\ k = 1..n)$ are iid gamma$(\theta)$ distributed, together with classical properties of gamma random variables.

Given the $k-$sample fell in $(S_1,..,S_p)$, consider then the random $p-$partition of unity $\mathbf{\Sigma}_p = (\Sigma_1,..,\Sigma_p)$, so with $\sum_1^p \Sigma_q = 1$.

For any subsequence $\{m_1 \neq m_2 \neq ... \neq m_p\}$ of $\{1,..,p\}$ the joint conditional probability that the $k-$sample visited $\{m_1 \neq m_2 \neq ... \neq m_p\}$ in that order and that there are $k_q$ sample within each $\Sigma_{m_q}$, $q = 1,..,p$ is

$$\mathbf{P}\left(M_1^{'} = m_1,..,M_p^{'} = m_p;\mathfrak{C}_{n,k}(1) = k_1,..,\mathfrak{C}_{n,k}(p) = k_p;P_{n,k} = p \mid \mathbf{S}_n\right)$$

$$= \binom{n}{p}\frac{k!}{\prod_{q=1}^p k_q!}\left(\sum_1^p S_q\right)^k \prod_{q=1}^p \Sigma_{m_q}^{k_q} \prod_{q=1}^{p-1}\frac{\Sigma_{m_q}}{1-\sum_{l=1}^q \Sigma_{m_l}}\Sigma_{m_p}$$

$$= \binom{n}{p} \frac{k!}{\prod_{q=1}^{p} k_q!} \left( \sum_{1}^{p} S_q \right)^k \prod_{q=1}^{p-1} \frac{\Sigma_{m_q}^{k_q+1}}{1 - \sum_{l=1}^{q} \Sigma_{m_l}} \Sigma_{m_p}^{k_p+1}.$$

Here $M_1^{'}, .., M_p^{'}$ is the sequence of the $p$ first fragments' numbers obtained from the sampling process by avoiding the ones previously encountered within $\mathbf{\Sigma}_p$; these were defined in Subsection 3.2. The coefficient $\frac{k!}{\prod_{q=1}^{p} k_q!}$ is the standard number of ways the sample could have arisen. The term $\prod_{q=1}^{p-1} \frac{\Sigma_{m_q}}{1-\sum_{l=1}^{q} \Sigma_{m_l}} \Sigma_{m_p}$ is the probability that sequence $\{m_1, m_2, ..., m_p\}$ emerges in that order from the sampling process within $\mathbf{\Sigma}_p$; it results of Eq. (16). The term $\prod_{q=1}^{p} \Sigma_{m_q}^{k_q}$ arises from the fact that $\Sigma_{m_q}$ is visited $k_q$ times, $q = 1, .., p$. Summing over $\{m_1 \neq m_2 \neq ... \neq m_p\}$ and averaging over $\mathbf{\Sigma}_p$, Eq. (17) gives

$$\sum_{\{m_1 \neq .. \neq m_p\}} \mathbf{E} \left[ \prod_{q=1}^{p-1} \frac{\Sigma_{m_q}^{k_q+1}}{1 - \sum_{l=1}^{q} \Sigma_{m_l}} \Sigma_{m_p}^{k_p+1} \right] =$$

$$\prod_{q=1}^{p-1} \left\{ \frac{\Gamma\left(1 + (p-q+1)\,\theta\right)}{\Gamma\left(1+\theta\right)\Gamma\left((p-q)\,\theta\right)} \frac{\Gamma\left(1+\theta+k_q\right)\Gamma\left((p-q)\,\theta + k_{q+1} + .. + k_p\right)}{\Gamma\left(1 + (p-q+1)\,\theta + k_q + .. + k_p\right)} \right\}.$$

Putting all this together gives the announced result $(i)$. The results $(ii)$ and $(iii)$ are consequences of the expression of $\mathbb{P}_\theta\left(P_{n,k} = p\right)$ in terms of Bell polynomials as shown in part $(ii)$ of Theorem 9. $\square$

Consider now the $k-$sample and let $m_1 \neq m_2 \neq ... \neq m_p$ denote the ordered number of the first, second,..., $p^{th}$ distinct animals sampled from $\mathbf{S}_n$ when only $P_{n,k} = p$ distinct fragments were visited. Let $\mathcal{C}_{n,k}\left(q\right)$, $q = 1, .., p$ be the number of animals of $q^{th}$ species to appear. We have

**Theorem 13** *For any $k_q \geq 1$, $q = 1, .., p$ satisfying $\sum_1^p k_q = k$ and any $p = 1, .., n \wedge k$, it holds*

$$\mathbb{P}_\theta \left( \mathcal{C}_{n,k} \left( 1 \right) = k_1, .., \mathcal{C}_{n,k} \left( p \right) = k_p; P_{n,k} = p \right) \tag{47}$$

$$= \frac{(k-1)!}{\prod_{q=1}^{p-1} \left( k - \sum_1^q k_i \right)} \frac{\Gamma \left( 1 + (n-p+1)\, \theta \right) \Gamma \left( \theta + k_p \right)}{\Gamma \left( 1 + \theta \right) \Gamma \left( (n-p+1)\, \theta + k_p \right) \Gamma \left( k_p \right)} \times \tag{48}$$

$$\prod_{q=1}^{p-1} \left\{ \frac{\Gamma \left( 1 + (n-q+1)\, \theta \right) \Gamma \left( \theta + k_q \right)}{\Gamma \left( 1 + \theta \right) \Gamma \left( (n-q)\, \theta \right) \Gamma \left( k_q \right)} \frac{\Gamma \left( (n-q)\, \theta + k_{q+1} + .. + k_p \right)}{\Gamma \left( (n-q+1)\, \theta + k_q + .. + k_p \right)} \right\}.$$

**Proof:** Given $\mathbf{S}_n$, the probability of the event in Eq. (47) is

$$\prod_{q=1}^{p} \binom{k - \sum_{i=1}^{q-1} k_i - 1}{k_q - 1} \times \sum_{m_1 \neq .. \neq m_p} \prod_{q=1}^{p} S_{m_q}^{k_q}$$

where $\{m_1 \neq .. \neq m_p\}$ are realizations of the SBP ordered sample $M_1^{'}, .., M_p^{'}$. From this and Eq. (14), we get

$$\mathbb{P}_\theta \left( \mathcal{C}_{n,k} \left( 1 \right) = k_1, .., \mathcal{C}_{n,k} \left( p \right) = k_p; P_{n,k} = p \right)$$

$$= \prod_{q=1}^{p} \binom{k - \sum_{i=1}^{q-1} k_i - 1}{k_q - 1} \times \mathbf{E} \left[ \prod_{q=1}^{p-1} \left\{ L_q^{k_q - 1} \left( 1 - \sum_{i=1}^{q} L_i \right) \right\} \times L_p^{k_p - 1} \right].$$

Observing from Eq. (10) that $1 - \sum_{i=1}^{q} L_i = \prod_{i=1}^{q} \overline{V}_i$ and using the independence of the $V_i$s, we get

$$Q \left( k_1, .., k_p \right) \quad : \quad = \mathbf{E} \left[ \prod_{q=1}^{p-1} \left\{ L_q^{k_q - 1} \left( 1 - \sum_{i=1}^{q} L_i \right) \right\} \times L_p^{k_p - 1} \right] \tag{49}$$

$$= \prod_{q=1}^{p-1} \mathbf{E} \left[ V_q^{k_q - 1} \overline{V}_q^{k_{q+1} + .. + k_p} \right] \times \mathbf{E} \left[ V_p^{k_p - 1} \right]. \tag{50}$$

Recalling $V_k \overset{d}{\sim} \text{beta}(1+\theta, (n-k)\,\theta)$, $\overline{V}_k \overset{d}{\sim} \text{beta}((n-k)\,\theta, 1+\theta)$, $k = 1, .., n-1$, using the same argument which was used in Theorem 3, this last probability

term $Q(k_1,..,k_p)$ reads

$$\prod_{q=1}^{p-1}\left\{\frac{\Gamma(1+(n-q+1)\theta)\Gamma(\theta+k_q)}{\Gamma(1+\theta)\Gamma((n-q)\theta)}\frac{\Gamma((n-q)\theta+k_{q+1}+..+k_p)}{\Gamma((n-q+1)\theta+k_q+..+k_p)}\right\}\times$$
$$\frac{\Gamma(1+(n-p+1)\theta)\Gamma(\theta+k_p)}{\Gamma(1+\theta)\Gamma((n-p+1)\theta+k_p)}.$$

Recalling $\sum_1^p k_q = k$, the number of arrangements term $\prod_{q=1}^p \binom{k-\sum_{i=1}^{q-1}k_i-1}{k_q-1}$ is also $\frac{(k-1)!}{\prod_{q=1}^{p-1}(k-\sum_1^q k_i)}\frac{1}{\prod_{q=1}^p\Gamma(k_q)}$ and the full result follows from regrouping terms.

□

Remark (the law of succession):

$(i)$ Consider Eq (47) and, with $m_r \in \{m_1,..,m_p\}$, let

$$\mathbf{P}(M_{k+1}=m_r \mid \mathcal{C}_{n,k}(1)=k_1,..\mathcal{C}_{n,k}(p)=k_p; P_{n,k}=p)$$

be the conditional probability that the $(k+1)^{th}$ sample is one from the previously encountered fragment already visited $k_r$ times. To evaluate this probability, the term $Q(k_1,..,k_p)$ in Eq. (49) has to be replaced by

$$Q(k_1,..,k_{r+1,..},k_p)=\mathbf{E}\left[\prod_{q=1:q\neq r}^{p-1}\left\{L_q^{k_q-1}\left(1-\sum_{i=1}^q L_i\right)\right\}L_r^{k_r}\left(1-\sum_{i=1}^r L_i\right)L_p^{k_p-1}\right]$$

substituting $k_r+1$ to $k_r$ in Eq. (49). The correcting term is found to be

$$\begin{aligned}&\frac{Q(k_1,..,k_{r+1,..},k_p)}{Q(k_1,..,k_p)}\\ =&\prod_{q=1}^{r-1}\left\{\frac{(n-q)\theta+k_{q+1}+..+k_p}{(n-q+1)\theta+k_q+..+k_p}\right\}\frac{\theta+k_r}{(n-r+1)\theta+k_r+..+k_p}\\ =&\frac{\theta+k_r}{n\theta+k}.\end{aligned}$$

This shows that, as for the ESF

$$\mathbf{P}_\theta\left(M_{k+1} \in \text{ species seen } k_r \text{ times} \mid \mathcal{C}_{n,k}(1)=k_1,..\mathcal{C}_{n,k}(p)=k_p; P_{n,k}=p\right)$$

$$=\frac{\theta+k_r}{n\theta+k},$$

which is again independent of $k_q$, $q\in\{1,..,p\}\setminus\{r\}$ and also of $p$.

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

$$\mathbf{P}_\theta\left(M_{k+1} \text{ is new} \mid \mathcal{C}_{n,k}(1)=k_1,..\mathcal{C}_{n,k}(p)=k_p; P_{n,k}=p\right)=\frac{(n-p)\,\theta}{n\theta+k},$$

which is independent of cell occupancies $k_1,..,k_p$ but depends on the number $p$ of distinct fragments already visited by the $k-$sample. Note that $p\leq(n-1)\wedge k$ if this probability is to be strictly positive. □

## The Kingman limit

Consider the situation where $n\uparrow\infty$, $\theta\downarrow 0$ while $n\theta=\gamma>0$. We give below a new proof of the celebrated Donnelly-Tavaré-Griffiths sampling formula (as from [25], page 10). Indeed, we have

**Corollary 14** *In the Kingman limit, the probabilities (42) and (47) both converge to*

$$\mathbb{P}^*_\gamma\left(\mathcal{C}_k(1)=k_1,..,\mathcal{C}_k(p)=k_p; P_k=p\right)=\frac{k!\gamma^p}{(\gamma)_k\prod_{q=1}^{p}(k_q+..+k_p)}. \qquad (51)$$

**Proof:** $(i)$ Consider first the probability displayed in (42). First, we have $\binom{n}{p} \sim \frac{n^p}{p!}$, $\frac{\Gamma(n\theta)\Gamma(p\theta+k)}{\Gamma(n\theta+k)\Gamma(p\theta)} \sim \frac{(k-1)!p\theta}{(\gamma)_k}$ . Next,

$$\prod_{q=1}^{p-1}\left\{\frac{\Gamma\left(1+(p-q+1)\,\theta\right)\Gamma\left(1+\theta+k_q\right)}{\Gamma\left(1+\theta\right)\Gamma\left((p-q)\,\theta\right)}\frac{\Gamma\left((p-q)\,\theta+k_{q+1}+..+k_p\right)}{\Gamma\left(1+(p-q+1)\,\theta+k_q+..+k_p\right)}\right\}\sim$$

$$\prod_{q=1}^{p-1}\left\{k_q!\,(p-q)\,\theta\right\}\prod_{q=1}^{p-1}\left\{\frac{\Gamma\left(k_{q+1}+..+k_p\right)}{\Gamma\left(1+k_q+..+k_p\right)}\right\}=$$

$$\theta^{p-1}\,(p-1)!\prod_{q=1}^{p-1}k_q!\times\Gamma\left(k_p\right)\prod_{q=2}^{p-1}\left\{\frac{1}{k_q+..+k_p}\right\}\frac{1}{(k_1+..+k_p)!}=$$

$$\theta^{p-1}\,(p-1)!\prod_{q=1}^{p}k_q!\times\prod_{q=2}^{p}\left\{\frac{1}{k_q+..+k_p}\right\}\frac{1}{(k_1+..+k_p)!}=$$

$$\theta^{p-1}\,(p-1)!\prod_{q=1}^{p}k_q!\times\prod_{q=1}^{p}\frac{1}{k_q+..+k_p}\times\frac{1}{(k-1)!}.$$

Multiplying this expression by the factor

$$\binom{n}{p}\frac{k!}{\prod_{q=1}^{p}k_q!}\frac{\Gamma\left(n\theta\right)\Gamma\left(p\theta+k\right)}{\Gamma\left(n\theta+k\right)\Gamma\left(p\theta\right)}\sim\frac{n^p}{p!}\frac{k!}{\prod_{q=1}^{p}k_q!}\frac{(k-1)!p\theta}{(\gamma)_k}$$

gives the result, recalling $n\theta=\gamma$.

$(ii)$ Consider next the probability displayed in (47). In the Kingman limit, we have

$$\begin{aligned}
&\mathbb{P}_\theta\left(\mathcal{C}_{n,k}\left(1\right)=k_1,..,\mathcal{C}_{n,k}\left(p\right)=k_p;P_{n,k}=p\right)\\
\sim\;&\frac{(k-1)!}{\prod_{q=1}^{p-1}\left(k-\sum_1^q k_i\right)}\frac{\Gamma\left(1+\gamma\right)}{\Gamma\left(\gamma+k_p\right)}\prod_{q=1}^{p-1}\left\{\frac{\Gamma\left(1+\gamma\right)}{\Gamma\left(\gamma\right)}\frac{\Gamma\left(\gamma+k_{q+1}+..+k_p\right)}{\Gamma\left(\gamma+k_q+..+k_p\right)}\right\}\\
=\;&\frac{\gamma^p\,(k-1)!}{\prod_{q=1}^{p-1}\left(k-\sum_1^q k_i\right)}\frac{\Gamma\left(\gamma\right)}{\Gamma\left(\gamma+k\right)}=\frac{\gamma^p k!}{(\gamma)_k}\frac{1}{\prod_{q=1}^{p}\left(k_q+..+k_p\right)},
\end{aligned}$$

observing that terms in the product cancel pairwise and recalling $\sum_1^p k_q=k$.

These results, with a different proof, can be found in [25], page 10. □

Summing over $\{k_1, .., k_p\}$ satisfying $k_q \geq 1$, $q = 1, .., p$ and $\sum k_q = k$ gives the limiting probability $\mathbb{P}^*_\gamma (P_k = p)$ that there are $p \leq k$ distinct species visited in the $k-$sample. We get

**Corollary 15** *With $s_{k,p}$ the absolute value of the first kind Stirling numbers, it holds that*

$$\mathbb{P}^*_\gamma (P_k = p) = \frac{\gamma^p s_{k,p}}{(\gamma)_k},\ p = 1, .., k \tag{52}$$

*and*

$$\mathbb{P}^*_\gamma (\mathcal{C}_k (1) = k_1, .., \mathcal{C}_k (p) = k_p \mid P_k = p) = \frac{k!}{s_{k,p} \prod_{q=1}^p (k_q + .. + k_p)}. \tag{53}$$

**Proof:** Another expression of $s_{k,p} := [\gamma^p] (\gamma)_k$ (as the coefficient of $\gamma^p$ in $(\gamma)_k$) is

$$s_{k,p} = k! \sum \frac{1}{\prod_{q=1}^p (k_q + .. + k_p)}$$

where the summation runs over the integers $k_q \geq 1$, $q = 1, .., p$ satisfying $\sum k_q = k$ (see [25], Appendix 2 page 18). □

Acknowledgments: I should like to thank Professor Warren J. Ewens, from the University of Pennsylvania, for stimulating discussions about his results on sampling theory in population genetics and the related literature. These largely motivated the present work.

# References


[1] *International Conference on Random Mappings, Partitions and Permutations*. Advances in Applied Probability, **1992**, 24, 761-777.

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

[3] Donnelly, P. The heaps process, libraries and size-biased permutation. Journal of Applied Probability, **1991**, 28, 321-335.

[4] Kingman, J.F.C. Random partitions in population genetics. Proceedings of the Royal Society. London. Series A, **1978**, 361, No 1704, 1–20.

[5] Ewens, W.J. Population genetics theory - the past and the future. In *Mathematical and Statistical Developments of Evolutionary Theory*, S. Lessard Edt., Kluwer, Dordrecht, **1990**.

[6] Huillet, T. Sampling problems for randomly broken sticks. Journal of Physics A: Math. & Gen., **2003**, 36, No 14, 3947-3960.

[7] Barrera, J.; Huillet, T.; Paroissin, C. Size-biased permutation of Dirichlet partitions and search-cost distribution. To appear in Probability in the Engineering & Informational Sciences, **2005**, No 1.

[8] Engen, S. *Stochastic abundance models*. Monographs on Applied Probability and Statistics, Chapman and Hall, London, **1978**.

[9] Kingman, J.F.C. *Poisson processes*. Clarendon Press, Oxford, **1993**.

[10] Patil, G.P.; Taillie, C. Diversity as a concept and its implications for random environments. Bulletin de l'Institut International de Statistique, **1977**, 4, 497-515.

[11] Donnelly, P. Partition structures, Pòlya urns, the Ewens sampling formula and the age of alleles. Theoretical Population Biology, **1986**, 30, 271-288.

[12] Flajolet, P.; Gardy, D.; Thimonier, L. Birthday paradox, coupon collectors, caching algorithms and self-organizing search. Discrete Applied Mathematics, **1992**, 39, 207-229.

[13] Tavaré, S.; Ewens, W.J. Multivariate Ewens distribution. Chapter 41 in *Discrete Multivariate Distributions*, N.L. Johnson, S. Kotz and N. Balakrishnan Edts, Wiley, New York, **1997**, 232-246.

[14] Pitman, J. Random discrete distributions invariant under size-biased permutation. Advances in Applied Probability, **1996**, 28, 525-539.

[15] Pitman, J. Coalescents with multiple collisions. Annals of Probability, **1999**, 27, No 4, 1870–1902.

[16] Pitman, J. Poisson-Dirichlet and GEM invariant distributions for split-and-merge transformation of an interval partition. Combinatorics, Probability and Computing, **2002**, 11, No 5, 501–514.

[17] Pitman, J.; Yor, M. The two parameter Poisson-Dirichlet distribution derived from a stable subordinator. Annals of Probability, **1997**, 25, 855-900.

[18] Sibuya, M.; Yamato, H. Ordered and unordered random partitions of an integer and the GEM distribution. Statistics & Probability Letters, **1995**, 25, No 2, 177-183.

[19] Huillet, T.; Martinez, S. Sampling from finite random partitions. Methodology and Computing in Applied Probability, **2003**, 5, No 4, 467-492.

[20] Ewens, W.J. Some remarks on the law of succession. Athens Conference on Applied Probability and Time Series Analysis (1995), Vol. **I**, 229–244, *Lecture Notes in Statistics*, Springer, New York, **1996**, No 114.

[21] Ewens, W.J. The sampling theory of selectively neutral alleles. Theoretical Population Biology, **1972**, 3**,** 87-112.

[22] Blackwell, D.; MacQueen, J.B. Ferguson distributions via Pólya urn schemes. The Annals of Statistics, **1973**, 1, 353–355.

[23] Holst, L. The Poisson-Dirichlet distribution and its relatives revisited. Preprint of the Royal Institute of Technology, Stockholm, Sweden, **2001**.

[24] Comtet, L. *Analyse combinatoire*. Tomes 1 et 2. Presses Universitaires de France, Paris, **1970**.

[25] Donnelly, P.; Tavaré, S. The age of alleles and a coalescent. Advances in Applied Probability, **1986**, 18, 1-19.