Sampling problems for randomly broken sticks

2003 J. Phys. A: Math. Gen. 36 3947

(http://iopscience.iop.org/0305-4470/36/14/302)

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND GENERAL

J. Phys. A: Math. Gen. **36** (2003) 3947–3960 PII: S0305-4470(03)57741-1

Sampling problems for randomly broken sticks

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: huillet@ptm.u-cergy.fr

Received 17 December 2002, in final form 18 February 2003
Published 26 March 2003
Online at stacks.iop.org/JPhysA/36/3947

Abstract

Consider the random partitioning model of a population (represented by a stick of length 1) into n species (fragments) with identically distributed random weights (sizes). Upon ranking the fragments' weights according to ascending sizes, let $S_{m:n}$ be the size of the mth smallest fragment. Assume that some observer is sampling such populations as follows: drop at random k points (the sample size) onto this stick and record the corresponding numbers of visited fragments. We shall investigate the following sampling problems: (1) what is the sample size if the sampling is carried out until the first visit of the smallest fragment (size $S_{1:n}$)? (2) For a given sample size, have all the fragments of the stick been visited at least once or not? This question is related to Feller's random coupon collector problem. (3) 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? For this problem, the distribution of the size-biased permutation of the species' weights, as the sequence of their weights in their order of appearance is needed and studied.

PACS numbers: 02.50.−r, 87.23.Cc

1. Introduction

Random division models of a population into a (possibly large) number n of species, fragments or valleys with random weights or sizes have received considerable attention in various domains of applications.

In disordered systems' physics, it was first recognized as an important issue in [4], as a result of phase space (in iterated maps or spin glass models at thermal equilibrium) being typically broken into many valleys, or attraction basins, each with random weight (see also [1], chapter 4, participation ratios). Problems involving the breakdown or splitting of some item into random component parts or fragments, also appear in many other

fields of interest: for example, the composition of rocks into component compounds in geology (splitting of mineral grains or pebbles), the composition of biological populations into species or the random allocation of memory in computer sciences, but also models for storage and search, gene frequencies in population genetics [14] and biological diversity [16].

All these applications are concerned with randomly broken objects and random splitting (see also [9], pp 25, 30 for further motivations in physics involving collision processes and absorption of light through randomly distributed spheres). Considering the random weights of the various species must sum to 1, by normalization, the typical phase space of these models is the interval of unit length, randomly split in such a way that the fragments' masses, sizes or energies must sum to 1. The random structure of the population is then characterized by the ranked sequence of fragments' weights or sizes. This was observed in [4] (in the large n thermodynamic limit, i.e., with a denumerable number of fragments).

There are, of course, many ways to break the interval at random into n pieces and so we have to be more specific. We shall focus here on the simplest 'fair' statistical model for splitting the interval into a *finite* number n of fragments, obtained from the following well-known random fragmentation construction: throw at random $n-1$ points on a stick of unit length and consider the induced random division of this stick into n fragments with lengths the distance between consecutive points. Call $S_{m,n}$, $m=1,\ldots,n$, the length (weight) of the mth piece. Then, although the sizes $S_{m,n}$, $m=1,\ldots,n$ of each fragment all share the same distribution, the population structure turns out to be far from trivial as there are obviously fragments that are more or less long. We shall call $S_{m:n}$, $m=1,\ldots,n$ the ranked sequence of fragments' lengths according to $S_{1:n}\leqslant\cdots\leqslant S_{n:n}$. Results on this specific random partition model and on the ranking procedure that are needed in the following are briefly recalled in section 2.

Suppose some population is split at random in this way into species numbered from 1 to n with corresponding random weights. Assume some observer wishes to sample this species' population. In the sampling process, the larger the weight $S_{m,n}$ of a species, the more likely the observer is to meet the corresponding species m. Thus, the sampling process can naturally be modelled as follows: throw at random k points on the unit stick and carefully note the numbers of visited fragments. Then these numbers constitute the observer's observations and k is the sample size.

Various sampling problems then naturally arise, some of which can be formulated as follows. Does the k-sample contain the same fragment twice (or more) by some coincidence? Have all fragments been visited or are there any undiscovered ones left in the k-sample? For deterministic partitions of the stick, these questions are known in the statistical literature as Feller's 'birthday' and 'coupon collector' sampling problems (in an attempt to answer the following everyday-life questions: what is the number of students needed in a class so that two students have the same birthday or so that the students in this class were born every day of the year with $n=365$). It is also of interest to count the number of fragments in the k-sample with exactly r representatives (the so-called fragments' vector count), and ask what is known about its Ewens-sampling distribution (a problem first propounded by [7] in the context of Poisson–Dirichlet distributions [5])? When $r=0$, this quantity is of particular interest in the applications; for example, in the Poincaré bombing problem when k-bombs are thrown at random on n districts of a city with random sizes, it interprets as the number of unbombed districts in a scattered shot. In geology (biology), it is the number of unobserved compounds (species) in the sampling process of a finite random partition of rocks (populations). These important problems will be considered elsewhere in this partition context and related ones.

Other related sampling problems of interest which constitute the main body of this work are the following.

(1) What is the sample size till the first visit to the smallest fragment of the partition whose length is $S_{1:n}$? This problem is addressed in section 3, the result being proposition 1 where the distribution function of this random variable is given.
(2) In what order are *new* fragments being discovered and what is the time separating the discovery of consecutive new fragments until they have all been exhausted? One intuitively expects these times to be increasing while approaching complete exhaustion of the list. These problems are addressed in sections 5 and 6. Our main results are summarized in theorems 8 and 10 where the distributions of these random waiting times are computed. For the last fragment to be discovered on the list, results on the sample size needed to visit all fragments are required.

For this last question, the distribution of the size-biased permutation of the fragments' lengths is needed: indeed, to avoid revisiting many times the firstly encountered species, we must remove it from the population once it has first been met in the sampling process and this requires an estimation of its weight. This process is described in section 4.1 with results displayed in proposition 3 and corollary 4. Once this is done, renormalizing the weights of the remaining species, we are left with a population with $n-1$ species, the sampling of which will necessarily supply a so far undiscovered species in the next step. Its weight can itself be estimated and so on, renormalizing again, until the whole available population species have been visited. This iterative process is described in section 4.2. It leads to the size-biased permutation of the species weights as the sequence of proportions of species in their order of appearance in a process of random sampling from the population. Thus, not only the visiting order of the different species can be understood but also their weights. Needed results on the weights of the size-biased permutation under our hypothesis are summarized in theorems 5 and 6 of section 4.

2. Preliminaries: the lengths of the pieces of a stick broken at random

Results on the population structure presented in this section are standard and can be found for example in [17, 3].

Consider $n-1$ independent identically distributed uniform draws $(X_1, \ldots, X_{n-1})$ on the interval $[0, 1]$. Putting this random vector into order, let $(X_{1:n}, \ldots, X_{n-1:n})$ be the ordered version of $(X_1, \ldots, X_{n-1})$ meaning $0 \leqslant X_{1:n} \leqslant \cdots \leqslant X_{m:n} \leqslant \cdots \leqslant X_{n-1:n} \leqslant 1$. Define the spacings between consecutive values by $S_{m,n} := X_{m:n} - X_{m-1:n}$, $m = 2, \ldots, n-1$. With $S_{1,n} := X_{1:n}$ and $S_{n,n} := 1 - X_{n-1:n}$ defining spacings at the endpoints, we are left with a partition of the interval $[0, 1]$ into n fragments with lengths $\mathbf{S}_n := (S_{m,n}; m = 1, \ldots, n)$ satisfying $\sum_{m=1}^{n} S_{m,n} = 1$. These have common distribution function ([9], p 22)

$$\overline{F}_{S_{m,n}}(s) := \mathbf{P}(S_{m,n} > s) = (1-s)^{n-1} \qquad s \in (0, 1) \tag{1}$$

independent of m, and so $S_{m,n} \stackrel{d}{=} S_n$ (equality in distribution). Such spacings are thus identically distributed with common distribution that of a beta random variable with parameters 1 and $n-1$. We shall put $S_{m,n} \stackrel{d}{=} S_n \stackrel{d}{=} \text{beta}(1, n-1)$.

(Recall that a random variable, say $B_{\alpha,\beta}$, with $B_{\alpha,\beta} \stackrel{d}{=} \text{beta}(\alpha, \beta)$, has density function $f_{B_{\alpha,\beta}}(x) := \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$; $\alpha, \beta > 0$; $x \in [0, 1]$ and moment function $\mathbf{E}B^{\lambda}_{\alpha,\beta} = \frac{\Gamma(\alpha+\lambda)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\alpha+\beta+\lambda)}$, with $\Gamma(\alpha)$ the Euler gamma function.)

This particular partition model therefore is one of splitting equitably a unit stick into statistically equal parts.

The sequence $(S_{m:n}; m = 1, \ldots, n)$ is obtained by ordering the fragments' sizes $(S_{m,n}; m = 1, \ldots, n)$, i.e., with $S_{1:n} \leqslant \cdots \leqslant S_{n:n}$. With $x_+ := \max(x, 0)$, the distribution function of $S_{m:n}$ is known to be [19]

$$F_{S_{m:n}}(s) = \sum_{q=m}^{n} \sum_{p=n-q}^{n} \frac{(-1)^{p+q-n} n!}{(n-q)!(n-p)!(q+p-n)!} (1 - ps)_+^{n-1}.$$

From this, $F_{S_{n:n}}(s) := \mathbf{P}(S_{n:n} \leqslant s) = \sum_{p=0}^{n} (-1)^p \binom{n}{p} (1 - ps)_+^{n-1}$ and $\overline{F}_{S_{1:n}}(s) := 1 - F_{S_{1:n}}(s) = (1 - ns)_+^{n-1}$ are the largest and smallest fragment size distributions for this partition. (Note the singular character of $S_{n:n}$'s distribution as underlined in [4].) As a result, one may prove that $\frac{n}{\log n} S_{n:n} \stackrel{a.s.}{\to} 1$, $n^2 S_{1:n} \stackrel{d}{\to} \exp(1)$ showing (see [11], p 297) that $S_{n:n}$ and $S_{1:n}$ are of order $\frac{\log n}{n}$ and n^{-2}, respectively, as n becomes large.

Although the division of the stick is fair in this random partition model, a high statistical variability of the fragments' sizes emerges in the thermodynamic limit.

3. The sample size till the first visit to the smallest fragment

In this section, we first model the sampling sieve process which seems to be relevant to our purposes, in an attempt to formalize the sentence: 'the larger the weight $S_{m,n}$ of a species is, the more likely is the observer to meet the corresponding species m'. Then, the problem of the sample size needed till the first visit to the smallest fragment of the partition is investigated.

Let $\mathbf{S}_n$ be the random partition of the interval into n fragments discussed above. Let $k \geqslant 1$ and $(U_1, \ldots, U_k)$ be k independent and identically distributed (i.i.d.) uniform random sample throws on $[0, 1]$. Let then $(M_1, \ldots, M_k)$ be the i.i.d. corresponding fragment numbers (the observer's observations). Then, their common conditional and unconditional distributions are given by

$$\mathbf{P}(M = m \mid \mathbf{S}_n) = S_{m,n} \qquad m \in \{1, \ldots, n\} \tag{2}$$

and

$$\mathbb{P}(M = m) := \mathbf{E}[\mathbf{P}(M = m \mid \mathbf{S}_n)] = \mathbf{E} S_{m,n} = \frac{1}{n} \tag{3}$$

which is consistent with our requirements. With $\mathbf{I}(\cdot)$ the set-indicator function, let $\mathcal{B}_{n,k}(m, \mathbf{S}_n) = \sum_{l=1}^{k} \mathbf{I}(M_l = m \mid \mathbf{S}_n)$ count the random number of occurrences of fragment m in the k-sample conditionally given $\mathbf{S}_n$. With $\sum_{m=1}^{n} \mathcal{B}_{n,k}(m, \mathbf{S}_n) = k$, the random variable $\mathcal{B}_{n,k}(m, \mathbf{S}_n)$ has binomial distribution, i.e., $\mathcal{B}_{n,k}(m, \mathbf{S}_n) \stackrel{d}{=} \mathrm{bin}(S_{m,n}, k)$ for which $\mathbf{P}(\mathcal{B}_{n,k}(m, \mathbf{S}_n) = b) = \binom{k}{b} S_{m,n}^b (1 - S_{m,n})^{k-b}$, $b = 0, \ldots, k$.

Let us now come to the sampling problem: what is the sample size that is needed till the first visit to the smallest fragment of the partition?

With $\mathcal{B}_{n,k}(m) = \sum_{l=1}^{k} \mathbf{I}(M_l = m)$ the random number of occurrences of fragment m in the k-sample, let $K_n(m) := \inf(k : \mathcal{B}_{n,k}(m) = 1)$ be the waiting time till the first visit to fragment m . Clearly, conditionally given $\mathbf{S}_n$, we have

$$\mathbf{P}(K_n(m) > k \mid \mathbf{S}_n) = \mathbf{P}(\mathcal{B}_{n,k}(m, \mathbf{S}_n) = 0) \qquad k \geqslant 1.$$

As $\mathcal{B}_{n,k}(m, \mathbf{S}_n) \stackrel{d}{=} \mathrm{bin}(S_{m,n}, k)$, we have

$$\mathbf{P}(K_n(m) > k \mid \mathbf{S}_n) = (1 - S_{m,n})^k$$

and so $K_n(m)$ is geometrically distributed conditionally given $\mathbf{S}_n$. For our purpose, let $m_{1:n} = \underset{m=1,\dots,n}{\arg\min}(S_{m,n})$ be the number of fragments with smallest size. Then

$$\mathbf{P}(K_n(m_{1:n}) > k \mid \mathbf{S}_n) = (1 - S_{1:n})^k$$

is the conditional waiting time till the first visit to this fragment. Averaging over the partitions $\mathbf{S}_n$, we obtain

$$\mathbb{P}(K_n(m_{1:n}) > k) := \mathbf{EP}(K_n(m_{1:n}) > k \mid \mathbf{S}_n) = \mathbf{E}[(1 - S_{1:n})^k].$$

In the uniform partition, $\overline{F}_{S_{1:n}}(s) = (1 - ns)_+^{n-1}$, so $\mathbf{P}[(1 - S_{1:n})^k > s] = 1 - (1 - n(1 - s^{1/k}))_+^{n-1}$. As a result, with $k \geqslant 1$, we get

$$\begin{aligned}
\mathbb{P}(K_n(m_{1:n}) > k) &= 1 - \int_0^1 (1 - n(1 - s^{1/k}))_+^{n-1}\,\mathrm{d}s \\
&= 1 - \int_{(1-\frac{1}{n})^k}^1 (1 - n(1 - s^{1/k}))^{n-1}\,\mathrm{d}s \\
&= 1 - \frac{k}{n}\left(1 - \frac{1}{n}\right)^{k-1} \int_0^1 x^{n-1}\left(1 + \frac{x}{n-1}\right)^{k-1}\mathrm{d}x \\
&= 1 - \frac{k}{n}\left(1 - \frac{1}{n}\right)^{k-1} \sum_{j=0}^{k-1} \binom{k-1}{j} (n-1)^{-j}(n+j)^{-1}.
\end{aligned}$$

So, we proved

Proposition 1. *For a stick broken at random, the probability distribution of the waiting time till the first visit to the smallest fragment reads*

$$\mathbb{P}(K_n(m_{1:n}) \leqslant k) = \frac{k}{n}\left(1 - \frac{1}{n}\right)^{k-1} \sum_{j=0}^{k-1} \binom{k-1}{j} (n-1)^{-j}(n+j)^{-1} \qquad k \geqslant 1$$

with $\mathbb{P}(K_n(m_{1:n}) = 1) = \frac{1}{n^2} = \mathbf{E}S_{1:n}$.

4. Sampling and size-biased permutation of the fragments

Consider the problem of determining the order in which the various species will be discovered in the sampling process. In order to avoid revisiting many times the same species once it has been discovered, we would like 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 weight is needed. This is possible as is shown in section 4.1 for the first visited species. Once this is done, after some renormalization of the remaining species' weights, we are left with a population of $n - 1$ species, the sampling of which will necessarily supply a so far undiscovered species. Its weight can itself be estimated and so forth, renormalizing again, until the whole available population species have been visited. This process is described in section 4.2. In this way, not only the visiting order of the different species can be understood but also their weights. The purpose of this section is to describe the statistical structure of the size-biased picked species' weights obtained while avoiding those previously encountered. This, among other things, will prove useful to solve the second problem on the sample sizes needed to discover consecutive new fragments.

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

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

$$\mathbf{P}(L_n = S_{m,n} \mid \mathbf{S}_n) = S_{m,n}.$$

In this size-biased picking procedure, long intervals are favoured and one expects that $L_n \succeq S_n$ in the following stochastic ordering sense.

Definition 2. *The random variable L_n is said to be stochastically larger than S_n (and we put $L_n \succeq S_n$) if*

$$\overline{F}_{S_n}(s) \leqslant \overline{F}_{L_n}(s) \qquad \forall s \in [0, 1].$$

Let us check that the size-biased pick is stochastically larger than the typical fragment's length in the uniform spacings case.

4.1. The length of the first size-biased pick

From the size-biased picking construction, it follows [6] that for all non-negative test functions f on $[0, 1]$,

$$\begin{aligned}\mathbb{E}[f(L_n)/L_n] &= \mathbf{E}[\mathbf{E}[f(L_n)/L_n \mid \mathbf{S}_n]] \\ &= \mathbf{E}\left[\sum_{m=1}^{n} f(S_{m,n})/S_{m,n}\mathbf{P}(L_n = S_{m,n} \mid \mathbf{S}_n)\right] = \mathbf{E}\left[\sum_{m=1}^{n} f(S_{m,n})\right]. \qquad (4)\end{aligned}$$

Taking in particular $f(x) = x\mathbf{I}\,(x > s)$ in (4), we get

$$\overline{F}_{L_n}(s) = \mathbf{E}\sum_{m=1}^{n} S_{m,n}\mathbf{I}(S_{m,n} > s)$$

which is

$$\overline{F}_{L_n}(s) = \sum_{m=1}^{n}\int_s^1 t\,\mathrm{d}F_{S_{m,n}}(t) = n\int_s^1 t\,\mathrm{d}F_{S_n}(t). \qquad (5)$$

Proposition 3. *It holds that*

$$L_n \succeq S_n$$

Proof. The condition $\overline{F}_{S_n}(s) \leqslant \overline{F}_{L_n}(s)$ holds for all s in $[0, 1]$ because this is equivalent to saying that $\int_s^1 t\,\mathrm{d}F_{S_n}(t)/\overline{F}_{S_n}(s) \geqslant \mathbf{E}(S_n)$ which is always true because the left-hand side is the conditional expectation of S_n given $S_n > s$, certainly larger than $\mathbf{E}(S_n)$ itself. As $S_n \stackrel{d}{=} \mathrm{beta}(1, n-1)$, this can be checked directly. Indeed, we obtain

$$\overline{F}_{L_n}(s) = ((n-1)s + 1)(1-s)^{n-1} \qquad (6)$$

showing that $L_n \stackrel{d}{=} \mathrm{beta}(2, n-1)$, with $\mathbf{E}L_n = 2/(n+1)$. One may then check that: $L_n \succeq S_n$, observing that $\overline{F}_{L_n}(s) = ((n-1)s + 1)(1-s)^{n-1} \geqslant \overline{F}_{S_n}(s) = (1-s)^{n-1}, \forall s \in [0, 1]$. □

This apparent paradox (discussed in [9], pp 22–23 and worked out in [11], pp 294–95) may be understood by observing that in the size-biased picking procedure, long intervals are favoured. It constitutes the version on the interval of the standard waiting-time paradox on the half-line [13]. As a corollary, the following decomposition holds ([2]):

Corollary 4. *Let B_n be a Bernoulli random variable with parameter $\frac{1}{n}$ and U a uniform random variable on* $[0, 1]$*, independent of B_n . Then, with R_n a* $[0, 1]$*-valued random variable with distribution*

$$R_n \stackrel{d}{=} B_n + (1 - B_n)U$$

the following decomposition holds

$$R_n L_n \stackrel{d}{=} S_n$$

where R_n and L_n are independent.

Proof. Since $\mathbf{P}(B_n = 1) = 1/n$, we have $\mathbf{E}R_n^q = \frac{1}{n} + \left(1 - \frac{1}{n}\right)\frac{1}{1+q}$. Taking $f(x) = x^{q+1}$ in (4), the moment function of L_n reads $(q > -2)$

$$\mathbb{E}\left[L_n^q\right] = \mathbf{E}\left[\sum_{m=1}^{n} S_{m,n}^{q+1}\right] = n\mathbf{E}\left[S_n^{q+1}\right] = \frac{\Gamma(n+1)\Gamma(q+2)}{\Gamma(n+q+1)}$$

recalling that $\mathbf{E}\left[S_n^q\right] = \frac{\Gamma(n)\Gamma(q+1)}{\Gamma(n+q)}$ is the common moment function of $S_{m,n}$, $m = 1, \ldots, n$, with $\mathbf{E}S_n = 1/n$. So, $\mathbf{E}\left[S_n^q\right] = \frac{n+q}{n(q+1)}\mathbf{E}\left[L_n^q\right] = \mathbf{E}R_n^q\mathbf{E}\left[L_n^q\right]$. □

4.2. Size-biased permutation of the fragments

Consider the random partition $\mathbf{S}_n$. Let $L_{1,n} := L_n$ be the first size-biased pick (SBP) just discussed for the first randomly chosen fragment $M_1 := M$, so with $L_{1,n} := S_{M_1,n}$. A standard problem is to iterate the size-biased picking procedure, by avoiding the fragments already encountered: by doing so, a size-biased permutation of the fragments is obtained. We would like to study this process here as it will prove useful in the following.

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

$$\mathbf{S}_n := \mathbf{S}_n^{(0)} \to (L_{1,n}, S_{1,n}, \ldots, S_{M_1-1,n}, S_{M_1+1,n}, \ldots, S_{n,n})$$

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

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

a new random partition of the unit interval into $n - 1$ random fragments.

Given $L_{1,n} \stackrel{d}{=} \text{beta}(2, n-1)$, the conditional joint distribution of the remaining components of $\mathbf{S}_n$ is the same as that of $(1 - L_{1,n})\mathbf{S}_{n-1}^{(1)}$ where the $(n - 1)$-vector $\mathbf{S}_{n-1}^{(1)}$ has the distribution of a uniform random partition into $n - 1$ fragments. Pick next at random an interval in $\mathbf{S}_{n-1}^{(1)}$ and call $\mathcal{L}_{2,n}$ its length, now with distribution $\text{beta}(2, n - 2)$, and iterate until all fragments have been exhausted.

With $\mathcal{L}_{1,n} := L_{1,n}$, the length of the second SBP by avoiding the first reads

$$L_{2,n} = (1 - \mathcal{L}_{1,n})\mathcal{L}_{2,n}.$$

Iterating, the final SBP vector is $\mathbf{L}_n := (L_{1,n}, \ldots, L_{n,n})$.

From this construction, if $(\mathcal{L}_{1,n}, \ldots, \mathcal{L}_{n-1,n})$ is an independent vector with distribution $\mathcal{L}_{m,n} \stackrel{d}{=} \text{beta}(2, n - m)$, $m = 1, \ldots, n - 1$, then

$$\begin{aligned} L_{m,n} &= \prod_{k=1}^{m-1}(1 - \mathcal{L}_{k,n})\mathcal{L}_{m,n} \qquad m = 1, \ldots, n-1 \\ L_{n,n} &= 1 - \sum_{m=1}^{n-1} L_{m,n} = \prod_{m=1}^{n-1}(1 - \mathcal{L}_{k,n}) \end{aligned} \tag{7}$$

is the stick-breaking scheme construction of the size-biased pick vector. Note that $\overline{\mathcal{L}}_{k,n} := 1 - \mathcal{L}_{k,n} \stackrel{d}{=} \text{beta}(n-k, 2)$ and that $\mathcal{L}_{n,n}$ can be set to 1: $\mathcal{L}_{n,n} \equiv 1$. From this well-known construction (see [15], chapters 9, 9.6; [16, 5]), we obtain that the $L_{m,n}$, $m = 1, \ldots, n$ are arranged in stochastically decreasing order. More precisely,

Theorem 5

(i) The law of $L_{m,n}$, for $m = 1, \ldots, n$, is characterized by

$$\mathbb{E}L_{m,n}^{\lambda} = \prod_{k=1}^{m-1} \mathbf{E}\overline{\mathcal{L}}_{k,n}^{\lambda} \mathbf{E}\mathcal{L}_{m,n}^{\lambda} = \prod_{k=1}^{m-1} \frac{\Gamma(n-k+\lambda)\Gamma(n-k+2)}{\Gamma(n-k+2+\lambda)\Gamma(n-k)} \times \frac{\Gamma(2+\lambda)\Gamma(2+n-m)}{\Gamma(2+n-m+\lambda)}.$$

(ii) Let $B_{n-m+1,1} \stackrel{d}{=} \text{beta}(n-m+1, 1)$. Then
$L_{m,n} \stackrel{d}{=} B_{n-m+1,1} \times L_{m-1,n}$, $m = 2, \ldots, n$, with independent $B_{n-m+1,1}$ and $L_{m-1,n}$, $m = 2, \ldots, n$.

(iii) $L_{1,n} \succeq \cdots \succeq L_{m,n} \succeq \cdots \succeq L_{n,n}$.

Proof. (i) is a direct consequence of the construction, since $\overline{\mathcal{L}}_{k,n} := 1 - \mathcal{L}_{k,n} \stackrel{d}{=} \text{beta}(n-k, 2)$, $k = 1, \ldots, m-1$, $\mathcal{L}_{m,n} \stackrel{d}{=} \text{beta}(2, n-m)$ are mutually independent. Also, if $B_{\alpha,\beta}$ is a beta(α, β) random variable, $\alpha, \beta > 0$, then $\mathbf{E}B_{\alpha,\beta}^{\lambda} = \frac{\Gamma(\alpha+\lambda)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\alpha+\beta+\lambda)}$ is its moment function, the corresponding expression of $\mathbf{E}L_{m,n}^{\lambda}$ follows. In particular, it holds $\mathbf{E}L_{m,n} = 2\frac{(n-m+1)}{n(n+1)}$, with $\sum_{m=1}^{n} \mathbf{E}L_{m,n} = 1$.

(iii) being a consequence of (ii), it remains to prove (ii). This can easily be proved by recurrence, observing first that

$$\mathbb{E}\big[L_{2,n}^{\lambda}\big] = \mathbf{E}[(1-\mathcal{L}_{1,n})^{\lambda}]\mathbf{E}[(\mathcal{L}_{2,n})^{\lambda}] = \frac{n-1}{q+n-1}\mathbf{E}\big[L_{1,n}^{\lambda}\big]$$

and that $\frac{n-1}{q+n-1}$ is the moment function of $B_{n-1,1} \stackrel{d}{=} \text{beta}(n-1, 1)$. □

Let us now compute the joint distribution of $\mathbf{L}_n$.

To this end, let us first discuss the visiting order of the fragments in the SBP process. For any permutation $(m_1, \ldots, m_n)$ of $(1, \ldots, n)$, with $M_1, \ldots, M_p$, $p = 1, \ldots, n$, the first p distinct fragment numbers which have been visited in this SBP sampling process, we have

$$\mathbf{P}(M_1 = m_1, \ldots, M_p = m_p \mid \mathbf{S}_n) = \prod_{k=1}^{p-1} \frac{S_{m_k,n}}{1 - \sum_{l=1}^{k} S_{m_l,n}} S_{m_p,n}$$

so that

$$\mathbf{P}(M_{p+1} = m_{p+1} \mid \mathbf{S}_n, M_1 = m_1, \ldots, M_p = m_p) = \frac{S_{m_p,n}}{1 - \sum_{l=1}^{p} S_{m_l,n}} S_{m_{p+1},n}.$$

Clearly, with $(n)_p = n(n-1)\cdots(n-p+1)$, we get

$$\mathbb{P}(M_1 = m_1, \ldots, M_p = m_p) := \mathbf{E}\mathbf{P}(M_1 = m_1, \ldots, M_p = m_p \mid \mathbf{S}_n) = \frac{1}{(n)_p}$$

averaging over the partitions $\mathbf{S}_n$: the $S_{m,n}$ being identically distributed, all sub-sequences $(m_1, \ldots, m_p)$ of $(m_1, \ldots, m_n)$ are equiprobable. The distribution of $M_1, \ldots, M_p$ is known as the Bose–Einstein distribution. Similarly, looking at the fragment lengths in a full SBP procedure, we have

$$\mathbf{P}(L_{1,n} = S_{m_1,n}, \ldots, L_{n,n} = S_{m_n,n} \mid \mathbf{S}_n) = \prod_{k=1}^{n-1} \frac{S_{m_k,n}}{1 - \sum_{l=1}^{k} S_{m_l,n}} S_{m_n,n}. \tag{8}$$

Consider now the joint moment function of the random size-biased pick vector $\mathbf{L}_{n-1} := (L_{1,n}, \ldots, L_{n-1,n})$. We observe that from (7) and the independence of the $\mathcal{L}_{m,n}$

$$\mathbb{E}\prod_{m=1}^{n-1} L_{m,n}^{\lambda_m} = \mathbf{E}\prod_{m=1}^{n-1}\prod_{k=1}^{m-1} \overline{\mathcal{L}}_{k,n}^{\lambda_m}\mathcal{L}_{m,n}^{\lambda_m} = \prod_{m=1}^{n-1} \mathbf{E}\left[\mathcal{L}_{m,n}^{\lambda_m}\overline{\mathcal{L}}_{m,n}^{\lambda_{m+1}+\cdots+\lambda_{n-1}}\right] \tag{9}$$

with $\mathcal{L}_{m,n} \stackrel{d}{=} \text{beta}(2, n-m)$, $\overline{\mathcal{L}}_{m,n} \stackrel{d}{=} \text{beta}(n-m, 2)$, $m = 1, \ldots, n-1$. (By convention, if $m = n-1$, the exponent $\lambda_{m+1} + \cdots + \lambda_{n-1}$ in the last product expression is set to zero.)

Also, inverting the transformation given in (7), $\mathcal{L}_{m,n} = L_{m,n}\big/\big(1 - \sum_{k=1}^{m-1} L_{k,n}\big)$, $m = 1, \ldots, n-1$, the joint density of the $L_{m,n}$ can be found after some elementary algebra using the mutual independence of $\mathcal{L}_{m,n}$, $m = 1, \ldots, n-1$. Putting all these together, we obtain

Theorem 6

(i) The joint moment function of the SBP vector $\mathbf{L}_{n-1} = (L_{1,n}, \ldots, L_{n-1,n})$ *reads*

$$\mathbb{E}\prod_{m=1}^{n-1} L_{m,n}^{\lambda_m} = \prod_{m=1}^{n-1} \frac{\Gamma(n-m+2)\Gamma(2+\lambda_m)}{\Gamma(n-m)} \frac{\Gamma(n-m+\lambda_{m+1}+\cdots+\lambda_{n-1})}{\Gamma(n-m+2+\lambda_m+\cdots+\lambda_{n-1})}.$$

(ii) With $\overline{s}_m := \sum_{k=1}^{m} s_k$, *the joint density of* $\mathbf{L}_{n-1}$ *at* $(s_1, \ldots, s_{n-1}) \in [0,1]^{n-1}$ *satisfying* $\overline{s}_{n-1} < 1$, *is given by*

$$f_{L_{1,n},\ldots,L_{n-1,n}}(s_1, .., s_{n-1}) = \prod_{m=1}^{n-1} \frac{\Gamma(n-m+2)}{\Gamma(n-m)} \frac{s_m(1-\overline{s}_m)^{n-m-1}}{(1-\overline{s}_{m-1})^{n-m+1}}$$

Proof

(i) If a random variable $X \stackrel{d}{=} \text{beta}(\alpha, \beta)$, with $\overline{X} := 1 - X$, we have

$$\begin{aligned}\mathbf{E}[X^{\lambda_1}\overline{X}^{\lambda_2}] &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 x^{\alpha+\lambda_1-1}(1-x)^{\beta+\lambda_2-1}\,\mathrm{d}x \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+\lambda_1)\Gamma(\beta+\lambda_2)}{\Gamma(\alpha+\beta+\lambda_1+\lambda_2)}.\end{aligned}$$

Adapting this computation, as $\mathcal{L}_{m,n} \stackrel{d}{=} \text{beta}(2, n-m)$, $\mathbf{E}\big[\mathcal{L}_{m,n}^{\lambda_m}\overline{\mathcal{L}}_{m,n}^{\lambda_{m+1}+\cdots+\lambda_{n-1}}\big]$ has the expression displayed inside the product from (9).

(ii) is a direct calculation of the image measure, the Jacobian of the inverse transformation $L_{.,n} \to \mathcal{L}_{.,n}$ being $1\big/\prod_{m=1}^{n-2}(1-\overline{s}_m)$. In this expression, $\overline{s}_m := \sum_{k=1}^{m} s_k$, $\overline{s}_0 := 0$. □

Last, we shall underline an implicit combinatorial identity that will prove useful in the following.

Lemma 7. *For any* $p \in \{1, \ldots, n-1\}$

$$\sum_{m_1 \neq \cdots \neq m_p} \mathbf{E}\left\{\prod_{k=1}^{p-1} \frac{S_{m_k,n}^{\lambda_k+1}}{1-\sum_{l=1}^{k} S_{m_l,n}} S_{m_p,n}^{\lambda_p+1}\right\}$$
$$= \prod_{k=1}^{p} \frac{\Gamma(n-k+2)\,\Gamma(2+\lambda_k)}{\Gamma(n-k)} \frac{\Gamma(n-k+\lambda_{k+1}+\cdots+\lambda_p)}{\Gamma(n-k+2+\lambda_k+\cdots+\lambda_p)}.$$

Proof. From (8), with $p \in \{1, \ldots, n-1\}$

$$\mathbf{P}\left(L_{1,n} = S_{m_1,n}, \ldots, L_{p,n} = S_{m_p,n} \mid \mathbf{S}_n\right) = \prod_{k=1}^{p-1} \frac{S_{m_k,n}}{1-\sum_{l=1}^{k} S_{m_l,n}} S_{m_p,n}$$

and so

$$\mathbb{E}\left\{\prod_{k=1}^{p} L_{k,n}^{\lambda_k}\right\} = \sum_{m_1\neq\cdots\neq m_p} \mathbf{E}\left\{\prod_{k=1}^{p-1} \frac{S_{m_k,n}^{\lambda_k+1}}{1-\sum_{l=1}^{k} S_{m_l,n}} S_{m_p,n}^{\lambda_p+1}\right\}.$$

From (i) of the last theorem 6, this expectation has the displayed expression. □

5. The times separating the discovery of new fragments in the SBP process

In this section, we study the times (sample sizes) needed to discover consecutive new fragments until they have all been visited. Intuitively, one expects these times to be increasing while approaching complete exhaustion in the sampling process; it seems indeed harder to discover a new species when many of them are already known simply because the larger the number of already observed species is, the larger the probability of visiting one of these (and no other). The previous considerations on the SBP process will prove useful to achieve this goal. However, the visiting time to the last fragment deserves special attention and special tools, discussed in section 6.

Let K_{p+1} be the time separating the discovery of the $(p+1)$th new fragment from the pth in the SBP process. Clearly, $K_1 = 1$ as the first throw necessarily leads to the discovery of a new fragment.

Next, with $k_{p+1} \geqslant 1$ and any $m_{p+1} \neq (m_1 \neq \cdots \neq m_p)$ themselves all distinct, we have

$$\mathbf{P}\left(K_{p+1} = k_{p+1}, M_{p+1} = m_{p+1} \mid M_1 = m_1, \ldots, M_p = m_p, \mathbf{S}_n\right) = \left[\sum_{q=1}^{p} S_{m_q}\right]^{k_{p+1}-1} S_{m_{p+1}}$$

as the event under consideration is realized when $k_{p+1}-1$ trials fail whereas the last trial is a success. As a result, summing over m_{p+1}

$$\mathbf{P}(K_{p+1} = k_{p+1} \mid M_1 = m_1, \ldots, M_p = m_p, \mathbf{S}_n)$$
$$= \sum_{m_{p+1}\neq(m_1\neq\cdots\neq m_p)} \left[\sum_{q=1}^{p} S_{m_q}\right]^{k_{p+1}-1} S_{m_{p+1}} = \left[\sum_{q=1}^{p} S_{m_q}\right]^{k_{p+1}-1} \left(1-\sum_{q=1}^{p} S_{m_q}\right)$$

which is a geometric distribution. Stated differently,

$$\mathbf{P}(K_{p+1} > k_{p+1} \mid M_1 = m_1, \ldots, M_p = m_p, \mathbf{S}_n) = \left[\sum_{q=1}^{p} S_{m_q}\right]^{k_{p+1}}.$$

Averaging over $m_1 \neq \cdots \neq m_p$,

$$\mathbf{P}(K_{p+1} > k_{p+1} \mid \mathbf{S}_n) = \sum_{m_1\neq\cdots\neq m_p} \left[\sum_{q=1}^{p} S_{m_q}\right]^{k_{p+1}} \mathbf{P}(M_1 = m_1, \ldots, M_p = m_p \mid \mathbf{S}_n)$$

where $\mathbf{P}(M_1 = m_1, \ldots, M_p = m_p \mid \mathbf{S}_n)$ is given from the previous section by

$$\mathbf{P}(M_1 = m_1, \ldots, M_p = m_p \mid \mathbf{S}_n) = \prod_{k=1}^{p-1} \frac{S_{m_k,n}}{1-\sum_{l=1}^{k} S_{m_l,n}} S_{m_p,n}.$$

Averaging over the partitions, we find

$$\mathbb{P}(K_{p+1} > k_{p+1}) = \mathbf{E}\mathbf{P}(K_{p+1} > k_{p+1} \mid \mathbf{S}_n)$$
$$= \sum_{m_1\neq\cdots\neq m_p} \mathbf{E}\left\{\left[\sum_{q=1}^{p} S_{m_q}\right]^{k_{p+1}} \prod_{k=1}^{p-1} \frac{S_{m_k,n}}{1-\sum_{l=1}^{k} S_{m_l,n}} S_{m_p,n}\right\} \tag{10}$$

with, in particular

$$\mathbb{E}(K_{p+1}) = \sum_{k_{p+1}\geqslant 0} \mathbb{P}(K_{p+1} > k_{p+1})$$
$$= \sum_{m_1\neq\cdots\neq m_p} \mathbf{E}\left\{\frac{1}{1-\sum_{q=1}^{p} S_{m_q}} \prod_{k=1}^{p-1} \frac{S_{m_k,n}}{1-\sum_{l=1}^{k} S_{m_l,n}} S_{m_p,n}\right\}$$
$$= \sum_{m_1\neq\cdots\neq m_p} \mathbf{E}\left\{\prod_{k=1}^{p} \frac{S_{m_k,n}}{1-\sum_{l=1}^{k} S_{m_l,n}}\right\}.$$

We would like to compute the distribution $\mathbb{P}(K_{p+1} > k)$. From (10) and the multinomial identity, we have

$$\mathbb{P}(K_{p+1} > k_{p+1}) = \sum_{\substack{l_1,\ldots,l_p\geqslant 0\\ \sum_1^p l_k = k_{p+1}}} \frac{k_{p+1}!}{\prod_{k=1}^{p} l_k!} \sum_{m_1\neq\cdots\neq m_p} \mathbf{E}\left\{\prod_{k=1}^{p} S_{m_k,n}^{l_k} \prod_{k=1}^{p-1} \frac{S_{m_k,n}}{1-\sum_{l=1}^{k} S_{m_l,n}} S_{m_p,n}\right\}$$
$$= \sum_{\substack{l_1,\ldots,l_p\geqslant 0\\ \sum_1^p l_k = k_{p+1}}} \frac{k_{p+1}!}{\prod_{k=1}^{p} l_k!} \sum_{m_1\neq\cdots\neq m_p} \mathbf{E}\left\{\prod_{k=1}^{p-1} \frac{S_{m_k,n}^{l_k+1}}{1-\sum_{l=1}^{k} S_{m_l,n}} S_{m_p,n}^{l_p+1}\right\}. \tag{11}$$

Using lemma 7, we finally obtain the following expressions:

Theorem 8

(i) For $p = 0, \ldots, n-2$, the distribution of K_{p+1} is given by

$$\mathbb{P}(K_{p+1} > k_{p+1}) = \sum_{\substack{l_1,\ldots,l_p\geqslant 0\\ \sum_1^p l_k = k_{p+1}}} \frac{k_{p+1}!}{\prod_{k=1}^{p} l_k!} \prod_{k=1}^{p} \frac{\Gamma(n-k+2)\Gamma(2+l_k)}{\Gamma(n-k)} \frac{\Gamma(n-k+l_{k+1}+\cdots+l_p)}{\Gamma(n-k+2+l_k+\cdots+l_p)}.$$

(ii) The mean values of K_{p+1} increase with p and are given by

$$\mathbb{E}(K_{p+1}) = \sum_{m_1\neq\cdots\neq m_p} \mathbf{E}\left\{\prod_{k=1}^{p} \frac{S_{m_k,n}}{1-\sum_{l=1}^{k} S_{m_l,n}}\right\} = \frac{n(n-1)}{(n-p)(n-p-1)}$$

$p = 0, \ldots, n-2$.

Proof

(i) is a direct application of lemma 7 starting from (11).

(ii) follows also from this lemma. At $\lambda_1 = \cdots = \lambda_{p-1} = 0$ and $\lambda_p = -1$, we have for all $p \in \{1, \ldots, n-1\}$

$$\sum_{m_1\neq\cdots\neq m_p} \mathbf{E}\left\{\prod_{k=1}^{p-1} \frac{S_{m_k,n}}{1-\sum_{l=1}^{k} S_{m_l,n}}\right\} = (n-p+1) \sum_{m_1\neq\cdots\neq m_{p-1}} \mathbf{E}\left\{\prod_{k=1}^{p-1} \frac{S_{m_k,n}}{1-\sum_{l=1}^{k} S_{m_l,n}}\right\}$$
$$= \prod_{k=1}^{p-1} \frac{\Gamma(n-k+2)}{\Gamma(n-k)} \frac{\Gamma(n-k-1)}{\Gamma(n-k+1)} \frac{\Gamma(n-p+2)}{\Gamma(n-p+1)}$$
$$= (n-p+1) \prod_{k=1}^{p-1} \frac{n-k+1}{n-k-1} = (n-p+1) \frac{n(n-1)}{(n-p+1)(n-p)}.$$

So, with $p \in \{1, \ldots, n-1\}$, we find

$$\mathbb{E}(K_p) = \sum_{m_1 \neq \cdots \neq m_{p-1}} \mathbf{E}\left\{ \prod_{k=1}^{p-1} \frac{S_{m_k,n}}{1 - \sum_{l=1}^{k} S_{m_l,n}} \right\} = \frac{n(n-1)}{(n-p+1)(n-p)}$$

and we can check $\mathbb{E}(K_{p+1}) > \mathbb{E}(K_p)$. □

6. The waiting time till the visit to the last fragment

There remains to compute the distribution of K_n which is not supplied by the last theorem 8. From (iii) of theorem 5, the last interval that remains to be discovered in the SBP process has a length which is stochastically smallest.

To our purpose, let us note that the following decomposition holds

$$1 + \sum_{p=1}^{n-1} K_{p+1} = K_n^+ \tag{12}$$

where

$$K_n^+ := \inf(k : \forall m, \mathcal{B}_{n,k}(m) \geqslant 1) \geqslant n$$

is the sample size needed until all fragments have been visited at least once. This variable is the one of interest in a random version of Feller's coupon collector problem (see [8], p 48). Note also that if the event $K_n^+ \leqslant k$ is realized, there are no fragments left undiscovered in a k-sample of the interval partition $\mathbf{S}_n$.

From Poisson embedding techniques ([12] and the references therein) or from combinatorial considerations [10], elementary computations show that

$$\mathbf{E}\big(K_n^+ \,\big|\, \mathbf{S}_n\big) = \int_0^\infty \left(1 - \prod_{m=1}^{n} (1 - \mathrm{e}^{-S_{m,n}t})\right) \mathrm{d}t. \tag{13}$$

Since $\mathbf{E}\big(K_n^+ \,\big|\, \mathbf{S}_n\big) = \sum_{k \geqslant 0} \mathbf{P}\big(K_n^+ > k \mid \mathbf{S}_n\big)$, we get the following general expression for the tail probability due to Flajolet *et al* and Holst (see [10], theorem 2, p 9, [12]).

Theorem 9. *Conditionally given* $\mathbf{S}_n$*, the distribution of the coupon collector waiting time with unequal and random probabilities* $S_{m,n}$ *is given by*

$$\mathbf{P}\big(K_n^+ \leqslant k \mid \mathbf{S}_n\big) = k![t^k] \prod_{m=1}^{n} \left[\sum_{q \geqslant 1} \frac{(S_{m,n}t)^q}{q!} \right]$$

where $[t^k]\varphi(t)$ *is the coefficient of* t^k *in the power series expansion of* $\varphi(t)$*.*

Averaging over the partition $\mathbf{S}_n$, we obtain

$$\mathbb{P}\big(K_n^+ \leqslant k\big) = \mathbf{E}\mathbf{P}\big(K_n^+ \leqslant k \mid \mathbf{S}_n\big)$$

which is the distribution of the random coupon collector problem arising in our random partition model when the probabilities themselves are random.

Theorem 10. *Consider the randomly broken stick model* $\mathbf{S}_n$*. Then, with* $k \geqslant n$

$$\mathbb{P}\big(K_n^+ > k\big) = \frac{\Gamma(n)}{\Gamma(n+k)} \sum_{l=1}^{n-1} (-1)^{l-1} \binom{n}{l} \frac{\Gamma((n-l)+k)}{\Gamma(n-l)} \tag{14}$$

and

$$\mathbb{E}K_n^+ = \sum_{k \geqslant n} \mathbb{P}\big(K_n^+ > k\big).$$

Proof. We will use the combinatorial identity

$$\prod_{m=1}^{n}(1+x_m) = 1 + \sum_{l=1}^{n} \sum_{1 \leqslant m_1 < \cdots < m_l \leqslant n} \prod_{j=1}^{l} x_{m_j}.$$

Putting $\sum_{j=1}^{l} S_{m_j,n} = 1 - \sum_{j \neq (1,\ldots,l)} S_{m_j,n}$, using the above identity, we find

$$k![t^k] \prod_{m=1}^{n} [\mathrm{e}^{S_{m,n}t} - 1] = 1 + \sum_{l=1}^{n-1} (-1)^l \sum_{1 \leqslant m_1 < \cdots < m_l \leqslant n} \left(\sum_{j \neq (1,\ldots,l)} S_{m_j,n} \right)^k .$$

Recalling that $\mathbf{P}\left(K_n^+ \leqslant k \mid \mathbf{S}_n\right) = k![t^k] \prod_{m=1}^{n} [\mathrm{e}^{S_{m,n}t} - 1]$ and noting that $f(\mathbf{S}_n) := \left(\sum_{j \neq (1,\ldots,l)} S_{m_j,n} \right)^k$ is a homogeneous function of the $S_{m,n}$ of degree $d = k$, (14) follows from Steutel's result (see [18], theorem 2, p 237) which states that $\mathbf{E} f(S_{1,n}, \ldots, S_{n,n}) = \frac{\Gamma(n)}{\Gamma(n+d)} \mathbf{E} f(T_1, \ldots, T_n)$, where $T_1, \ldots, T_n$ are i.i.d. variables with exponential distribution of mean 1. □

From theorem 10, (12) and (ii) of theorem 8, the expectation $\mathbb{E}K_n$ of the time separating the discovery of the $(n-1)$th fragment from the last nth follows directly.

7. Conclusion and perspectives

The following simple sampling problems from finitely randomly broken sticks have been considered and solved: what is the sample size till the first visit to the smallest fragment of the partition? Given a sample size, have all fragments been visited or are there any still remaining to be discovered? In what order are new fragments being discovered and how long should one wait between the discovery of consecutive new fragments until the list is exhausted? Although these problems are easy to formulate, the answers turn out to be surprisingly difficult to derive. The required tools to achieve our task have been introduced and applied to yield new results. Essentially, some information on the size-biased permutation of the species' weights is required.

It would be interesting to ask the same questions for other stick-breaking models, and to consider the case of a denumerable number of fragments. However, from the sampling point of view, some questions asked appear meaningless in this enlarged context: for example, asking for the number of unvisited fragments in a k-sample is an absurd question, together with the one on the sample size needed to visit the smallest fragment. Rather, in that case, the way the number of visited fragments grows with the sample size seems to be the right question to ask. We are currently investigating these kinds of problems.

References

[1] Baldassarri A 1999 Statistique des évènements extrèmes et persistants *PhD Thesis* SPEC, CEA Saclay, Université Paris 11

[2] Collet P, Huillet T and Martinez S 2002 Finite random partitions of the interval *Preprint* (2002 *J. Appl. Prob.* at press)

[3] Darling D A 1953 On a class of problems related to the random division of an interval *Ann. Math. Stat.* **24** 239–53

[4] Derrida B and Flyvbjerg H 1987 Statistical properties of randomly broken objects and of multivalley structures in disordered systems *J. Phys. A: Math. Gen.* **20** 5273–88
[5] Donnelly P 1986 Partition structures, Pòlya urns, the Ewens sampling formula and the age of alleles *Theor. Pop. Biol.* **30** 271–88
[6] Engen S 1978 *Stochastic Abundance Models (Monographs on Applied Probability and Statistics)* (London: Chapman and Hall)
[7] Ewens W J 1972 The sampling theory of selectively neutral alleles *Theor. Pop. Biol.* **3** 87–112
Ewens W J 1972 *Theor. Pop. Biol.* **3** 240
Ewens W J 1972 *Theor. Pop. Biol.* **3** 376
[8] Feller W 1968 *An Introduction to Probability Theory and Its Applications* 3rd edn vol 1 (New York: Wiley)
[9] Feller W 1971 *An Introduction to Probability Theory and Its Applications* 2nd edn vol 2 (New York: Wiley)
[10] Flajolet P, Gardy D and Thimonier L 1992 Birthday paradox, coupon collectors, caching algorithms and self-organizing search *Disc. Appl. Math.* **39** 207–29
[11] Hawkes J 1981 On the asymptotic behaviour of sample spacings *Math. Proc. Camb. Phil. Soc.* **90** 293–303
[12] Holst L 1995 The general birthday problem *Random Struct. Algorithms* **6** 201–8
[13] Huillet T 2002 On the waiting time paradox and related topics *Fractals* **10** 173–88
[14] Kingman J F C 1978 Random partitions in population genetics *Proc. R. Soc.* A **361** 1–20
[15] Kingman J F C 1993 *Poisson Processes* (Oxford: Clarendon)
[16] Patil G P and Taillie C 1977 Diversity as a concept and its implications for random environments *Bull. Int. Stat. Inst.* **4** 497–515
[17] Pyke R 1965 Spacings (with discussion) *J. R. Stat. Soc.* B **27** 395–449
[18] Steutel F W 1967 Random division of an interval *Statistica Neerlandica* **21** 231–44
[19] Stevens W L 1939 Solution to a geometrical problem in probability *Ann. Eugenics* **9** 315–20