Ann Inst Stat Math (2014) 66:833–864
DOI 10.1007/s10463-013-0427-8

Prediction in Ewens–Pitman sampling formula and random samples from number partitions

Masaaki Sibuya

Received: 31 October 2011 / Revised: 11 June 2013 / Published online: 2 October 2013

Abstract Motivated by marine ecological data on species abundance, with the record of subsamples, two problems are investigated in this paper, assuming the Ewens–Pitman sampling formula: One is the prediction of the number of new species if the catch is continued, and the other is how the number of species will decrease in random subsamples. Related statistics and extended models are also considered. A tool for the work is the generalized Stirling numbers of three variables.

Keywords Bell polynomials · Gibbs partitions · Partition data · Pólya's urn model · Random number partitions · Random sum models · Size index · Trawl fishery · Waiting time.

1 Introduction

1.1 Random partition data

The basic data of ecological surveys on species abundance are the count, c_i, of individuals of the ith species, $i = 1, \dots k$, where k is the number of different species, and $n = c_1 + \cdots + c_k$ is the total number of observed individuals. We assume that species of zero count is neglected and that the order of the species is irrelevant. That is, the observation is a partition of n to a sum of k positive integers. Let $\mathscr{P}_{n,k}$ denote the set

Electronic supplementary material The online version of this article (doi:10.1007/s10463-013-0427-8) contains supplementary material, which is available to authorized users.

M. Sibuya (✉)
Professor Emeritus, Keio University, Hiyoshi, Kohoku-Ku,
Yokohama, Kanagawa 223-8522, Japan
e-mail: sibuyam@1986.jukuin.keio.ac.jp

of all such partitions and $\mathscr{P}_n := \bigcup_{k=1}^{n} \mathscr{P}_{n,k}$. For modeling species abundance data, it is natural to regard them as a realization of the random partition, a probability measure on $\mathscr{P}_n$. Charalambides (2007) serves as an introduction to random partitions. A strategy for modeling partition data is discussed by Hoshino (2012). A celebrated random partition is the Ewens–Pitman sampling formula (EPSF), developed in population genetics, and thoroughly and profoundly investigated by Pitman (2006).

Assume that an EPSF partition in $\mathscr{P}_{m,k}$ is observed, and the observation on $\mathscr{P}_{m+n}$, $m = 1, 2, \ldots$ is continued. The conditional process started from $\mathscr{P}_{m,k}$ will be called restart EPSF. Statistics of the restart process are investigated, and, for example, the moments of the number of new species are shown. In the framework of the nonparametric Bayesian statistics (Ferguson 1973), our prediction is the nonparametric estimation in the consistent Gibbs random partition, and some basic results are shown in Lijoi et al. (2007, 2008). EPSF is a typical consistent Gibbs process, and the distributions of its statistics are expressed in closed forms. Further, EPSF belongs to the random sum models of the Gibbs partitions, which are not consistent, except for EPSF, and less flexible. However, more known models are available.

Subsampling in species abundance survey In ecological surveys of fish or small insects, thousands of individuals are caught, and a part of the catch is randomly selected and species of all individuals in the subsample are identified. See, e.g., Heales et al. (2000, 2003a,b) and van Ark and Meiswinkel (1992). Researchers are anxious about the possible inhomogeneity of subsampling and larger fluctuation of subsamples.

The reverse process of the restart EPSF process is the simple random sampling without replacement from the conditional distribution on $\mathscr{P}_{n,k}$. In the consistent Gibbs partition, individuals are sequentially numbered, and deletion from the partition of total size n is just to delete n, but we have to see where is n. Here, the backward equations of the Gibbs (Markov) process and of the restart process are shown. More basically, simple random sampling without replacement from a partition sample is investigated, and the moments of the number of species in subsamples are obtained.

Contents of the paper In Sect. 1, EPSF, as a balls-to-urns process, is introduced, and for later use the generalized Stirling numbers are introduced. In Sect. 2.1, the restart of the balls-to-urns process is defined, and the p.d.f. and moments of the number of non-empty urns in the restart process are shown. In Sect. 2.2, the random partitions in the restart process are investigated, and in Sect. 2.3, waiting times are treated. Section 3 states briefly two basic approaches to random partitions. Section 4 studies basic facts of the simple random subsamples of partition data (Sect. 4.1) and sampling from random partitions (Sect. 4.2). Based on Sects. 2 and 4, a dataset in Heales et al. (2003a) is analyzed in Sect. 5. In Sect. 6, the implication of the results of this paper are discussed.

1.2 Ewens–Pitman sampling formula

Partitions of a number Partitions $\{c_1, c_2, \ldots\} \in \mathscr{P}_n$ are expressed in several ways. The simplest one is descending order statistics (DOS), $c_1 \geq \cdots \geq c_k$. Another simple one is ascending order statistics, and to avoid duplicate numbers, a standard expression is

$$s=(s_1,\dots,s_n),\quad s_j:=\sum_{i=1}^{k}\mathbb{I}[c_i=j],\quad 1\le j\le n,$$

where $\sum_{j=1}^{n} s_j = k, \sum_{j=1}^{n} js_j = n$, and $\mathbb{I}[\cdot]$ is the predicate: $\mathbb{I}[\text{True}] = 1$, and $\mathbb{I}[\text{False}] = 0$. This expression is called *frequency of frequencies* by Good (1953) and *size index* by Sibuya (1993). Ecologists call it *species abundance distribution* (see, e.g., McGill et al. 2007). For other expressions, see Andrews and Eriksson (2004).

Balls-to-urns process EPSF is best illustrated by Pólya-type balls-to-urns process (Yamato and Sibuya 2003b). Suppose that Balls $(B_1, B_2, \dots)$ are put into Urns $(U_1, U_2, \dots)$ one by one at random as follows:

(i) First, B_1 is put into U_1 with the probability 1. (ii) At stage $n,\ n = 1, 2, \dots,$ assume $(B_1, \dots, B_n)$ are in $(U_1, \dots, U_k)$ such that there is no empty urn, and c_j balls in $U_j;\ 1 \le j \le k \le n,\ c_1 + \cdots + c_k = n$. Now, B_{n+1} is put into $U_j;\ 1 \le j \le k + 1$ with the probabilities

$$\frac{c_j-\alpha}{\theta+n},\ \text{if } 1\le j\le k;\quad \frac{\theta+k\alpha}{\theta+n},\ \text{if } j=k+1.$$

The result, at stage n, is a random partition A of the index set $[n] := \{1, \dots, n\}$, with the restriction

$$1\in A_1;\ \min(\ell;\ell\in[n]\setminus\cup_{i=1}^{j}A_i)\in A_{j+1},\quad j=1,\dots,k-1.$$

The probability of a partition A of $[n],\ n \in \mathbb{N}$ is, by induction of the model,

$$\begin{aligned} p((c_1,\dots,c_k))\ &:= \mathbb{P}(A) = \frac{1}{(\theta|-1)_n}\prod_{j=1}^{k}(\theta+(j-1)\alpha)(1-\alpha|-1)_{c_j-1},\\ c_j\ &:= |A_j|, 1\le j\le k, c_j>0,\ c_1+\cdots+c_k=n, \end{aligned}\tag{1}$$

where r.v. k is the number of subsets of A, $(a|b)_n := a(a-b)\dots(a-(n-1)b)$. The set of (θ, α), for which $\mathbb{P}(A) \ge 0,\ \forall A$, partition of $[n], n = 1, 2, \dots,$ is

$$0\le\alpha\le 1, -\alpha\le\theta,\quad \text{or } \alpha<0, \theta=-M\alpha,\ M=1,2,\dots$$

Ewens–Pitman sampling formula If both the balls and the urns are indistinguishable, a sequence of random partitions on $\mathscr{P}_n,\ n = 1, 2, \dots,$ is, in terms of the size index, as follows:

$$\begin{aligned} w(n;s) :=\ \mathbb{P}\{S_n=s\} &= \frac{(\theta|-\alpha)_k}{(\theta|-1)_n}\pi_n(s)\prod_{j=1}^{n}((1-\alpha|-1)_{j-1})^{s_j},\\ \pi_n(s)\ &= \frac{n!}{\prod_{j=1}^{n}s_j!(j!)^{s_j}},\quad \text{if } s=(s_1,\dots,s_n)\in\mathscr{P}_{n,k}. \end{aligned}\tag{2}$$

Random partitions (2) are called *Ewens–Pitman sampling formula* and denoted by EPSF$(n; \theta, \alpha)$. Their most important property is *the partition structure*:

$$\begin{aligned} &w(n; (s_1+1, s_2, \ldots, s_n))\frac{s_1+1}{n} \\ &\quad + \sum_{j=2}^{n} w(n; (s_1, \ldots, s_{j-1}-1, s_j+1, \ldots, s_n))\frac{(s_j+1)j}{n}\mathbb{I}[s_{j-1}>0] \\ &= w(n-1; (s_1, \ldots, s_{n-1})). \end{aligned} \tag{3}$$

Equation (3) means that if one ball is randomly deleted from a random partition EPSF $(n; \theta, \alpha)$, a random partition EPSF$(n-1; \theta, \alpha)$ is obtained. On the parameter estimation of EPSF, see Carlton (1999), Sibuya and Yamato (2001), and Hoshino (2001).

Number of non-empty urns. Suppose S_n is EPSF$(n; \theta, \alpha)$ partition, and put $K_n := \sum_{j=1}^{n} S_j$, the number of non-empty urns, or the number of species. Its probability mass function (p.m.f.) $f_n(k) := \mathbb{P}\{K_n = k\}$ satisfies,

$$f_{n+1}(k) = \frac{n-k\alpha}{\theta+n} f_n(k) + \frac{\theta+(k-1)\alpha}{\theta+n} f_n(k-1). \tag{4}$$

The p.m.f. $f_n(k)$ is expressed in a closed form,

$$f_n(k) = \frac{(\theta|-\alpha)_k}{(\theta|-1)_n} S_{n,k}(-1, -\alpha, 0), \tag{5}$$

where $S_{n,k}(a, b, c)$ is the generalized Stirling number introduced below. The distribution (5) is denoted by EPSF-K(θ, α).

If $K_n = k$, or if $S \in \mathbb{P}_{n,h}$, the conditional distribution of random partitions is

$$\mathbb{P}\{S = s | S \in \mathscr{P}_{n,k}\} = \frac{1}{S_{n,k}(-1, -\alpha, 0)} \pi_n(s) \prod_{j=1}^{n} ((1-\alpha|-1)_{j-1})^{s_j}, \tag{6}$$

which is independent of θ.

Generalized Stirling numbers For use in Sect. 2, recall 3-parameter generalized Stirling numbers (G3SN), defined by the polynomial identity in t:

$$(t+c|a)_n \equiv \sum_{k=0}^{n} S_{n,k}(a, b, c)(t|b)_k. \tag{7}$$

The properties of G3SN are summarized in the Appendix. In the next section, we need its convolution-type recurrence,

Proposition 1

$$S_{m+n,\ell}(a,b,c) = \sum_{j=0}^{\min(m,\ell)} S_{m,j}(a,b,c) S_{n,\ell-j}(a,b,c+jb-ma),$$
$$\forall \ell, m, n \in \mathbb{N},\ 0 \le \ell \le n. \tag{8}$$

Proof In the polynomial identity in t, $(t+c|a)_{m+n} = (t+c|a)_m (t+c-ma|a)_n$,

$$\text{RHS} = \sum_{k=0}^{m}\sum_{\ell=0}^{n} S_{m,k}(a,b,c)(t|b)_k S_{n,\ell}(a,b,c+kb-ma)(t-kb|b)_\ell.$$

Compare the coefficient of $(t|b)_k$ in both sides to obtain (8). □

2 Restart process

2.1 Number of non-empty urns in restart process

The number of parts or non-empty urns, K_n, is regarded as random walks on a square grid, $\{(n,k), 1 \le k \le n\}$, starting from (1,1) and moving from (n,k) to $(n+1,k)$ or $(n+1,k+1)$. In this subsection, we consider the process starting from any grid point, moving with the same transition probability as the original process.

Restarting from (m,k) refresh the count, that is, change the states from $(m,k),\dots,(m+n,k+\ell)$ to $(0,0),\dots,(n,\ell)$, $0 \le \ell \le n$, or we consider the random variable

$$K_n(m,k) := (K_{m+n} - K_m)|(K_m = k). \tag{9}$$

Its p.m.f. $f_n(\ell) := \mathbb{P}\{K_n(m,k) = \ell\}$ satisfies, from (4),

$$f_{n+1}(\ell) = \frac{m+n-(k+\ell)\alpha}{\theta+m+n} f_n(\ell) + \frac{\theta+(k+\ell-1)\alpha}{\theta+m+n} f_n(\ell-1),$$
$$0 \le \ell \le n+1,\ f_1(0) = 1. \tag{10}$$

In contrast to K_n, $K_n(m,k)$ can be 0 with positive probability, and it is seen that, from (10) without computing the p.m.f.,

$$f_n(0) = \frac{(m-k\alpha|-1)_n}{(\theta+m|-1)_n}.$$

It decreases fast when n increases, unless α and θ are small enough.

We generalize EPSF-K (5), using Proposition 1, comparing the forward Eqs. (4) and (10).

$$\mathbb{P}\{K_{m+n} = \ell\} = \frac{S_{m+n,\ell}(-1, -\alpha, 0)(\theta| - \alpha)_\ell}{(\theta| - 1)_{m+n}} = \sum_{k=0}^{\min(m,\ell)} \frac{S_{m,k}(-1, -\alpha, 0)(\theta| - \alpha)_k}{(\theta| - 1)_m} \times \frac{S_{n,\ell-k}(-1, -\alpha, m - k\alpha)(\theta + k\alpha| - \alpha)_{\ell-k}}{(\theta + m| - 1)_n}. \tag{11}$$

Hence, the p.m.f. of $K_n(m, k)$ of (9) is, replacing ℓ by $k + \ell$ in (11),

$$f_n(\ell) = \mathbb{P}\{K_n(m, k) = \ell\} = \frac{(\theta + k\alpha| - \alpha)_\ell}{(\theta + m| - 1)_n} S_{n,\ell}(-1, -\alpha, m - k\alpha), \quad 0 \le \ell \le n, \tag{12}$$

which will be denoted by RsEPSF-K(θ, α).

This p.m.f. is given by Lijoi et al. (2007), Equation 8. They treat the restart process in a more general framework, which will be sketched in Sect. 3.2. The expression $f_n(0)$ obtained below (10) is confirmed by $S_{n.0}(a, b, c) = (c|a)_n$.

Moments of restart process To obtain numerical values of moments in usual applications, it is practical to calculate $f_n(\ell)$ recursively by (10) and $\sum_\ell \ell^r f_n(\ell)$ naively. For theoretical purposes some expressions are necessary.

Proposition 2 *Moments of* $K_n = K_n(m, k)$ (Eq. 9) *are as follows:*

$$E((K_n)_r) = \frac{(\theta + k\alpha| - \alpha)_r}{\alpha^r(\theta + m| - 1)_n} \sum_{j=0}^{r} \binom{r}{j} (-1)^{r-j} (\theta + m + j\alpha| - 1)_n, \tag{13}$$

$$E((K_n| - 1)_r) = \frac{\theta + k\alpha}{\alpha^r(\theta + m| - 1)_n} \times \sum_{j=0}^{r} \binom{r}{j} (-1)^{r-j} (\theta + (k + j - 1)\alpha|\alpha)_{r-1} (\theta + m + j\alpha| - 1)_n. \tag{14}$$

A general expression (15) *is shown in the proof.*

Proof First, calculate $E(\theta + (k + K_n)\alpha| - \alpha)_r)$. Since

$$\begin{aligned}
&(\theta + k\alpha| - \alpha)_\ell(\theta + (k + \ell)\alpha| - \alpha)_r = (\theta + k\alpha| - \alpha)_r(\theta + (k + r)\alpha| - \alpha)_\ell,\\
&\sum_{\ell=0}^{n} f_n(\ell)(\theta + (k + \ell)\alpha| - \alpha)_r = \frac{(\theta| - \alpha)_r}{(\theta + m| - 1)_n} \cdot \sum_{\ell=0}^{n} S_{n,\ell}(-1, -\alpha, m - \alpha)(\theta + (k + r)\alpha| - \alpha)_\ell\\
&\quad = \frac{(\theta + k\alpha| - \alpha)_r}{(\theta + m| - 1)_n}(\theta + m + r\alpha| - 1)_n.
\end{aligned}$$

Next, to calculate $E((K_n|\epsilon)_r)$, use (7),

$$(\ell|\epsilon)_r = \sum_{j=0}^{r} S_{r,j}(\epsilon, -1, -c)(c+\ell|-1)_j,$$
$$c = \theta/\alpha + k, \quad (c+\ell|-1)_j = \alpha^{-j}(\theta + (k+\ell)\alpha|-\alpha)_j,$$
$$E((K_n|\epsilon)_r) = \sum_{j=0}^{r} S_{r,j}(\epsilon, -1, -c)\frac{(\theta+k\alpha|-\alpha)_j(\theta+m+j\alpha|-1)_n}{\alpha^j(\theta+m|-1)_n}. \quad (15)$$

For special cases, note that

$$S_{n,k}(1,-1,c) = \binom{n}{k}(c-k)_{n-k}, \quad S_{n,k}(-1,-1,-c) = \binom{n}{k}(c)_{n-k}(-1)^{n-k}.$$

□

Remarks (a) Typical examples are

$$E(K_n) = \frac{\theta+k\alpha}{\alpha}\left(\frac{(\theta+m+\alpha|-1)_n}{(\theta+m|-1)_n} - 1\right),$$
$$E((K_n)_2) = \frac{(\theta+k\alpha|-\alpha)_2}{\alpha^2(\theta+m|-1)_n} \times((\theta+m+2\alpha|-1)_n - 2(\theta+m+\alpha|-1)_n + (\theta+m|-1)_n).$$

Put $m = k = 0$ to obtain the moments of EPSF-K(θ, α), which is shown in Yamato and Sibuya (2003a) by induction.

(b) For Ewens sampling formula EPSF-K$(\theta, 0)$, factorial cumulants are

$$\kappa_{(r)} = (-1)^{r-1}(r-1)!\theta^r\sum_{j=0}^{n-1}\frac{1}{(\theta+j)^r} = \theta^r(\psi^{(r-1)}(\theta+n) - \psi^{(r-1)}(\theta)), \quad r = 1, 2, \ldots,$$

where $\psi^{(r)}(\theta)$ is the polygamma function, $\psi^{(0)}(\theta) = \psi(\theta)$ is the digamma function, and

$$E(K_n) = \theta(\psi(\theta+n) - \psi(\theta)) \approx \theta\left(\log(\theta+n) - \psi(\theta) - \frac{1}{2(n+\theta)}\right), \ n \to \infty.$$

(c) $S_{n,k}(0, b, c)$ or $S_{n,k}(a, 0, c)$, which appear in the case $\epsilon = 0$, are known as Carlitz's weighted Stirling numbers (Hsu and Shiue 1998).

2.2 Partition in restart process

In the previous subsection, the p.m.f. in RsEPSF-K and its moments are shown for the application of Sect. 5. In this subsection, random partition in the restart process RsEPSF is obtained based on the results of the previous subsection.

In our restart process, at the restart time $n = 0$, the old partition is neglected, but data (m, k), or $S_m \in \mathscr{P}_{m,k}$, are recorded. The new balls put into old urns are mixed and counted as $S_{n,0}$, the number of *virtual urns of size 0*. Hence, in the restart process, a number n is partitioned at random to the sum of *nonnegative* integers, and the new size index is such that

$$s = (s_0, s_1, \dots, s_n),\ \ s_j \geq 0,\ j = 0, 1, \dots, \quad n = s_0 + \sum_{j=1}^{n} j s_j,\ \ \ell := \sum_{j=1}^{n} s_j. \quad (16)$$

The p.m.f. of RsEPSF-K (12) is rewritten as follows, because of (31) in Appendix.

$$\begin{aligned}\mathbb{P}\{K_n = \ell\} &= \frac{(\theta + k\alpha|-\alpha)_\ell}{(\theta + m|-1)_n} S_{n,\ell}(-1, -\alpha, m - k\alpha)\\ &= \frac{(\theta + k\alpha|-\alpha)_\ell}{(\theta + m|-1)_n} \sum_{n^*=0}^{n} \binom{n}{n^*} (m - k\alpha|-1)_{n-n^*} S_{n^*,\ell}(-1, -\alpha, 0)\\ &= \sum_{n^*=0}^{n} \binom{n}{s_0} \frac{(m - k\alpha|-1)_{s_0} (\theta + k\alpha|-1)_{n^*}}{(\theta+m|-1)_n} \frac{(\theta+k\alpha|-\alpha)_\ell}{(\theta+k\alpha|-1)_{n^*}} S_{n^*,\ell}(-1, -\alpha, 0)\end{aligned} \quad (17)$$

where $n* = n - s_0$. In the last expression, the summand is the joint p.m.f. of (K_n, s_0):

$$\begin{aligned} f_n(\ell, s_0) &= h_n(s_0) f_n(\ell|s_0) = h_n(s_0) f_{n^*}(\ell),\\ f_n(\ell, s_0) &= \binom{n}{s_0} \frac{(\theta + k\alpha|-\alpha)_\ell (m - k\alpha|-1)_{s_0}}{(\theta + m|-1)_n} S_{n^*,\ell}(-1, -\alpha, 0), \quad n^* + s_0 = n,\\ h_n(j) &= \binom{n}{j} \frac{(m - k\alpha|-1)_j (\theta + k\alpha|-1)_{n-j}}{(\theta + m|-1)_n}, \quad 0 \leq j \leq n. \end{aligned} \quad (18)$$

The factor $f_{n^*}(\ell)$ is the conditional p.m.f. of $\sum_{j=1}^{n} s_j$, which is EPSF-K$(\theta + k\alpha, \alpha)$, see (5). The p.m.f. $h_n(j)$ is the negative hypergeometric distributions NgHg$(n; m - k\alpha, \theta + k\alpha)$ with (θ, α) in the EPSF parameter space.

Another derivation of the joint p.m.f. $f_n(\ell, s_0)$ is to consider the Markov process with the states (n, ℓ, s_0) and the forward equation,

$$\begin{aligned} f_{n+1}(\ell, s_0) = \frac{n^* - \ell\alpha}{\theta + m + n} f_n(\ell, s_0) + \frac{\theta + k\alpha + (\ell - 1)\alpha}{\theta + m + n} f_n(\ell - 1, s_0)\\ + \frac{m - k\alpha + s_0 - 1}{\theta + m + n} f_n(\ell, s_0 - 1), \end{aligned} \quad (19)$$

which generalizes (10). Starting from $f_1(0, 0) = 1$, $f_n(\ell, s_0)$ is shown by induction.

Given (K_n, s_0) the conditional random partition is (6), where n replaced by n^*, and

$$\mathbb{P}\{S_n = s\} = \sum_{j=0}^{n} \mathbb{P}\{S_n^+ = s^+ | S_{n,0} = j\}\mathbb{P}\{S_{n,0} = j\}, \quad S_n^+ := S_n \backslash S_{n,0},$$

which determine the p.m.f. of RsEPSF in the following proposition.

Proposition 3 (Lijoi et al. 2008) *In the restart process* RsEPSF$(m, k; \theta, \alpha)$*, random partition S_n has the following p.m.f., with the symbol s of* (16)*,*

$$\mathbb{P}\{S_n = s\}$$
$$= \frac{(\theta + k\alpha| - \alpha)_\ell (m - k\alpha| - 1)_{s_0}}{(\theta + m| - 1)_n} \binom{n}{s_0} \pi_{n^*}(s^+) \prod_{j=1}^{n^*} ((1-\alpha| - 1)_{j-1})^{s_j},$$
$$n^* = n - s_0, \quad s^+ := (s_1, \ldots, s_n). \tag{20}$$

Since the joint moments of the size index of EPSF are known, those for RsEPSF are simply their NgHg mixtures. For example,

$$E((S_{n,j})_r) = \frac{(n)_{jr}(\theta + k\alpha| - \alpha)(\theta + m + r\alpha| - 1)_{n-jr}}{(\theta + m| - 1)_n} \prod_{j=2}^{n} \left(\frac{(1-\alpha| - 1)_{j-1}}{j!} \right)^{r_j},$$
$$1 \le j \le n, \; r = 1, 2, \ldots$$
$$E(S_{n,1}) = \frac{n(\theta + k\alpha)(\theta + m + \alpha| - 1)_{n-1}}{(\theta + m| - 1)_n}.$$

The distribution of $S_{n,0}$ following (18), NgHg$(m - k\alpha, \theta + k\alpha)$ has the asymptotic property; $S_{n,0}/n \xrightarrow{d}$ Be$(m - k\alpha, \theta + k\alpha)$, the beta distribution, and the ratio of the number of balls in the old and new urns, is rather stable.

Proposition 3 is shown in Lijoi et al. (2008), Equation (22) as an example of the general theorem on the nonparametric Bayesian estimation of the consistent Gibbs partitions. The above statement that the conditional p.m.f. of $\sum_{j=1}^{n} s_j$ is EPSF-K$(\theta + k\alpha, \alpha)$ is shown, in Lijoi et al. (2008), to hold for any consistent Gibbs partitions.

A restriction on old species observation In RsEPSF, only the condition $K_m = k$ is assumed. Suppose that the counts $c = \{c_1, \ldots, c_k\}$ of individuals of species are also recorded, and, in the restart process, the counts of the new species and those of some specific ones of c are interested. That is, $c \in \mathscr{P}_{n,k}$ is divided into two groups, say, $\{c_1, \ldots, c_{k_1}\} \in \mathscr{P}_{m_1,k_1}$ and $\{c_{k_1+1}, \ldots, c_k\} \in \mathscr{P}_{m_2,k_2}$, $m_1 + m_2 = m$, $k_1 + k_2 = k$. Accordingly, s_0 is divided into $s_{01} + s_{02} = s_0$. A problem in RsEPSF is to find the probability of $(s_{02}, K_n(m, k)) = (0, \ell)$ at a stage n.

Given $s_0 = s_{01} + s_{02}$, $K_n(m, k)$ is independent of $(s_{01}, s_{02})|s_0$. Since s_0 follows the negative hypergeometric distribution (18), (s_{01}, s_{02}) follows the multivariate (actually bivariate) negative hypergeometric distribution MvGHg$(n; m_1 - k_1\alpha, m_2 - k_2\alpha, \theta + k\alpha)$. Hence, by (17),

$$\begin{aligned}
&\mathbb{P}\{(s_{02}, K_n(m,k)) = (0,\ell) | m_1 + m_2 = m \ \& \ k_1 + k_2 = k\} \\
&\quad = \frac{(\theta + k\alpha | -\alpha)_\ell}{(\theta + m | -1)_n} \sum_{s_0=0}^{n} \binom{n}{s_0} (m_1 - k_1\alpha | -1)_{s_0} S_{n^*,\ell}(-1, -\alpha, 0), \\
&\quad = \frac{(\theta + k\alpha | -\alpha)_\ell}{(\theta + m | -1)_n} S_{n,\ell}(-1, -\alpha, m_1 - k_1\alpha), \quad n^* = n - s_0, \ 0 \le s_0 \le n, \\
&\mathbb{P}\{s_{02} = 0 | m_1 + m_2 = m \ \& \ k_1 + k_2 = k\} \\
&\quad = \sum_{\ell=0}^{n} \frac{(\theta + k\alpha | -\alpha)_\ell}{(\theta + m | -1)_n} S_{n,\ell}(-1, -\alpha, m_1 - k_1\alpha) = \frac{(\theta + k_2\alpha + m_1 | -1)_n}{(\theta + m | -1)_n}.
\end{aligned}$$

These are shown in Lijoi et al. (2008) Sect. 3.3, *Looking backward*.

2.3 Waiting time

Write $K_n = K_n(m,k)$ for short, and regard (n, K_n) as random walks on the square lattice. Two types of waiting time are important in applications. One is W_ν, the first time as $K_n = \nu$, that is, the number of individuals to be observed to find ν new species. The other is V_μ, the first time as $n - K_n = \mu$, that is, the number of individuals of already found species reaches μ, which may be the time to give up the observation efforts.

First, we examine W_ν, and note that

$$W_\nu = n \quad \Leftrightarrow \quad K_{n-1} \ge \nu - 1 \ \& \ K_n \ge \nu \quad \Leftrightarrow \quad K_{n-1} = \nu - 1 \ \& \ K_n - K_{n-1} = 1,$$

where K_n has the p.m.f. (12). The probability $\mathbb{P}\{K_n - K_{n-1} = 1 | K_{n-1} = \nu - 1\}$ is seen in the second factor of (10).

Proposition 4 *In the restart process* RsEPSF-K(θ, α)*, the waiting time* W_n *for the first arrival of* $K_n(m,k)$ *to* ν *has the p.m.f.*

$$\mathbb{P}\{W_\nu = n\} = \frac{(\theta + k\alpha | -\alpha)_\nu}{(\theta + m | -1)_n} S_{n-1,\nu-1}(-1, -\alpha, m - k\alpha), \quad n \ge \nu \ge 1. \tag{21}$$

Remarks

(a) Since $S_{n,0}(a,b,c) = (c|a)_n$,

$$\mathbb{P}\{W_1 = n\} = (m + k\alpha | -1)_{n-1}(\theta + k\alpha)/(\theta + m | -1)_n,$$

which is consistent with $f_n(0)$ below (10).

(b) The sum of (21) over $\nu \le n < \infty$ is one, and in general, as a rational generating function,

$$\frac{1}{(t|b)_k} = \sum_{n=k}^{\infty} S_{n-1,k-1}(a,b,c) \frac{1}{(t + c|a)_n},$$

which was shown by Corcino (2001).

(c) For RsEPSF-K$(\theta, 0)$, the Ewens case,

$$\mathbb{P}\{W_\nu = n\} = \frac{\theta^\nu}{(\theta + m|-1)_n} S_{n-1,\nu-1}(-1, 0, m), \quad n \ge \nu \ge 1.$$

Sibuya and Nishimura (1997) applied this model to the prediction of waiting time for record-breaking, assuming m to be a known or unknown real parameter.

Moments of W_ν Since

$$\frac{(\theta+m+n-1)_r}{(\theta+m-r|-1)_r(\theta+m|-1)_n} = \frac{1}{(\theta+m-r|-1)_n}, \text{ where } (a)_r = (a|1)_r,$$

$$\sum_{n=\nu}^{\infty} S_{n-1,\nu-1}(-1,-\alpha,m-k\alpha)\frac{1}{(\theta+m-r|-1)_n} = \frac{1}{(\theta+k\alpha-r|-\alpha)_\nu},$$

$$E((W_\nu+\theta+m-1)_r) = (\theta+m-1)_r \frac{(\theta+k\alpha|-\alpha)_\nu}{(\theta+k\alpha-r|-\alpha)_\nu}, \quad r = 1, 2, \ldots,$$

which can be solved recursively. For example,

$$E(W_\nu) = (\theta+m-1)\left(\frac{(\theta+k\alpha|-\alpha)_\nu}{(\theta+k\alpha-1|-\alpha)_\nu} - 1\right),$$

$$E((W_\nu)_2) = (\theta+m-1)_2\left(\frac{(\theta+k\alpha|-\alpha)_\nu}{(\theta+k\alpha-2|-\alpha)_\nu} - 1\right)$$
$$-2(\theta+m-1)^2\left(\frac{(\theta+k\alpha|-\alpha)_\nu}{(\theta+k\alpha-1|-\alpha)_\nu} - 1\right).$$

If $\theta = r - k\alpha$, $E((W_\nu)_2) = \infty$, $r = 1, 2, \ldots$

Next, we examine V_μ. Similar to the case of W_ν,

$$V_\mu = n \quad \Leftrightarrow \quad K_{n-1} = n-\mu \;\&\; K_n = n-\mu \quad \Leftrightarrow \quad K_{n-1} = n-\mu \;\&\; K_n = K_{n-1}.$$

Hence,

$$\begin{aligned}
&\mathbb{P}\{V_\mu = n\}\\
&= \frac{(\theta+k\alpha|-\alpha)_{n-\mu}(m+n-1-(k+n-\mu)\alpha)}{(\theta+m|-1)_n} S_{n-1,n-\mu}(-1,-\alpha,m-k\alpha)\\
&= \left(\frac{(\theta+k\alpha|-\alpha)_{n-\mu}}{(\theta+m|-1)_{n-1}} - \frac{(\theta+k\alpha|-\alpha)_{n-\mu+1}}{(\theta+m|-1)_n}\right) S_{n-1,n-\mu}(-1,-\alpha,m-k\alpha),\\
&\quad n \ge \mu \ge 1.
\end{aligned} \tag{22}$$

For RsEPSF-K$(\theta, 0)$, the Ewens case,

$$\mathbb{P}\{V_\mu = n\} = \frac{\theta^{n-\mu}(m+n-1)}{(\theta+m|-1)_n} S_{n-1,n-\mu}(-1, 0, m).$$

The moments of V_μ are rather complicated even in the Ewens case.

3 Extensions of EPSF model

3.1 Random sum models

Let $Z_1, Z_2, \ldots,$ be an i.i.d. sequence of *positive* integer-valued random variables, and Y be a nonnegative integer-valued random variable, which is independent of Z_is. Then, the conditional distribution of the random vector

$$\{Z_1, \ldots, Z_Y\}|(X := Z_1 + \cdots + Z_Y = n),$$

is a random partition on $\mathscr{P}_n,\ n = 1, 2, \ldots$ Let $f(w)$, $g(w)$, $h(w)$ be the probability generating functions of X, Y and Z, respectively, *in the exponential form*:

$$\mathbb{P}\{Z = z\} = h_z.\ h_0 = 0,\quad h(w) = \sum_{n=0}^{\infty} \frac{n! h_n w^n}{n!} =: \sum_{n=0}^{\infty} \frac{\breve{h}_n w^n}{n!},\quad \text{or}\quad h_n = \left[\frac{w^n}{n!}\right] \frac{h(w)}{n!}.$$

Note that

$$f(w) = g(h(w)),\quad \left[\frac{w^n}{n!}\right] g(h(w) = \sum_{m=0}^{n} \mathrm{B}_{n,m}(\breve{h}) g_m,$$

$$\mathrm{B}_{n,m}(\phi) := \left[\frac{w^n}{n!}\right] \frac{\phi(w)^m}{m!} = \sum_{s \in \mathscr{P}_{n,m}} n! \prod_{j=1}^{n} \frac{1}{s_j!} \left(\frac{\phi_j}{j!}\right)^{s_j},\quad 1 \le m \le n,$$

$$\mathrm{B}_{n,0}(\phi) = \mathbb{I}[n = 0],\ \phi(w) = \sum_{n=0}^{\infty} \frac{\phi_n w^n}{n!},$$

where $\mathrm{B}_{n,m}(\phi)$ are homogeneous polynomials in $\phi_1, \ldots, \phi_{n-m+1}$ of the degree m, called *the partial exponential Bell polynomials*.

The random partition $\{Z_1, \ldots, Z_Y\}|(X = n)$, in terms of the size index, has the p.m.f.

$$\mathbb{P}\{S = s\} = \frac{g_m n!}{\mathrm{B}_n(\breve{g}; \breve{h})} \prod_{j=1}^{n} \frac{1}{s_j!} \left(\frac{\breve{h}_j}{j!}\right)^{s_j},\quad \text{if}\ S \in \mathscr{P}_{n,m},$$

where $\mathrm{B}_n(\breve{g}; \breve{h})$ is the normalizing constant. The number of parts $K_n := S_1 + \cdots + S_n$ has the p.m.f.

$$\mathbb{P}\{K_n = k\} = \frac{g_m}{\mathrm{B}_n(\breve{g}; \breve{h})} \mathrm{B}_{n,k}(\breve{h}),\quad 1 \le k \le n.$$

Under the condition $S \in \mathscr{P}_{n,m}$,

$$\mathbb{P}\{S = s | S \in \mathscr{P}_{n,m}\} = \frac{n!}{\mathrm{B}_{n,m}(\breve{h})} \prod_{j=1}^{n} \frac{1}{s_j!} \left(\frac{\breve{h}_j}{j!} \right)^{s_j}, \tag{23}$$

which is independent of (g_m).

Proposition 5 *Under the condition $S \in \mathscr{P}_{n,m}$ of the random sum partitions (23),*

$$E\left(\prod_{j=1}^{n} (S_j)_{q_j} | S \in \mathscr{P}_{n,m}\right) = (n)_Q \frac{\mathrm{B}_{n-Q,m-q}(\breve{h})}{\mathrm{B}_{n,m}(\breve{h})} \prod_{j=1}^{n} \left(\frac{\breve{h}_j}{j!} \right)^{q_j}, \quad q = \sum_{j=1}^{n} q_j, \quad Q = \sum_{j=1}^{n} j q_j,$$

Hence,

$$E(K_n | S \in \mathscr{P}_{n,m}) = \frac{1}{\mathrm{B}_{n,m}(\breve{h})} \sum_{k=1}^{n} \binom{n}{j} \breve{h}_j \mathrm{B}_{n-j,m-1}(\breve{h}).$$

There are many known *compound distributions* $f(w) = g(h(w))$, see, e.g., Johnson et al. (2005) and Charalambides (2005). However, the p.m.f. (23) is not always expressed in closed forms, and the discussions in previous sections cannot be developed in a general way. The Bell polynomials satisfy some recurrence formula, and they are at least numerically computable. See, e.g., Comtet (1974) and Charalambides (2002) on the Bell polynomials, and the Appendix for the relationship with G3SN.

It is shown that (Kerov 2006) if Z_i is the extended (or Engen) truncated negative binomial, the random partition is EPSF(θ, α), if Y is the negative binomial $(0 < \alpha < 1)$, the Poisson $(\alpha = 0)$, or the negative hypergeometric $(\alpha < 0,\ \theta = -M\alpha,\ M = 2, 3, \ldots)$ r.v.

3.2 Gibbs partition models

Another extension is possible. In expression (1), EPSF is regarded as an ordered partition $A = (A_1, \ldots, A_k)$ of the set $[n]$, where the probability $p(c_1, \ldots, c_k)$ is a symmetric function of $c_j = |A_j|, 1 \le j \le k \le n$. Such a p.m.f. on $\mathscr{P}_n$ is called exchangeable partition probability function (EPPF). This is a *size-biased-random-permutation* expression of EPSF without the contraction to the size index. The EPPF in the form

$$p((n_1, \ldots, n_k)) = V_{n,k} \prod_{j=1}^{k} W_{n_j},$$

where $V_{nk}(1 \le k \le n)$ and $W_n(n = 1, 2, \dots)$ are nonnegative sequences, called the *Gibbs form*. A Gibbs form is *consistent* if

$$p((n_1, \dots, n_k)) = \sum_{j=1}^{k} p((n_1, \dots, n_j + 1, \dots, n_k)) + p((n_1, \dots, n_k, 1)),$$

which is equivalent to the partition structure (3) of the random number partitions. It turns out that the Gibbs form is consistent iff

$$p((n_1, \dots, n_k)) = V_{n,k} \prod_{j=1}^{k} (1 - \alpha | -1)_{n_j - 1}, \quad (n_1, \dots, n_k) \in \mathscr{P}_{n,k}, -\infty < \alpha < 1, \tag{24}$$

where $V_{n,k}(1 \le k \le n)$ satisfies the backward recursion

$$V_{n,k} = (n - k\alpha) V_{n+1,k} + V_{n+1,k+1}. \tag{25}$$

For EPSF(θ, α), $V_{n,k} = (\theta | -\alpha)_k / (\theta | -1)_n$. The random sum models in Sect. 3.1 also have the Gibbs form with the weight $V_{n,k} = V_k / c_n$. However, they are not consistent except for the EPSF.

A property of the consistent Gibbs partition is that its conditional random partition on a specific $\mathscr{P}_{n,k}$ is given by (6), which is shown there for EPSF. See Gnedin and Pitman (2006), Equation (6). Conversely, the consistent Gibbs partitions are characterized by the mixing distribution, $\mathbb{P}\{K_n = k\}$. This property is essential as shown in Sect. 4.2.

Lijoi et al. (2008) show that our restart process can be defined for the random partitions of the Gibbs form and can calculate the distributions of some statistics. Among others, using the symbols of Sect. 2.1

$$g_n(\ell) := \mathbb{P}\{K_n(m, k) = \ell\} = \frac{V_{m+n,k+\ell}}{V_{m,k}} S_{n,\ell}(-1, -\alpha, m - k\alpha),$$

which leads immediately to (12). Following the line of the paper, its recurrence formula, corresponding to (10), is

$$g_{n+1}(k) = (m + n - (k + \ell)\alpha) \frac{V_{m+n+1,k+\ell}}{V_{m+n,k+\ell}} g_n(k) + \frac{V_{m+n+1,k+\ell}}{V_{m+n,k+\ell-1}} g_n(k - 1).$$

Similarly, extending the waiting time (21),

$$\mathbb{P}\{W_\nu = n\} = \frac{V_{m+n,k+\nu}}{V_{m,k}} S_{n-1,\nu-1}(-1, -\alpha, m - k\alpha), \quad n \ge \nu \ge 1.$$

For the basic work on the Gibbs form, see Gnedin and Pitman (2006), and for the related works, see references in Lijoi et al. (2008).

Within the broader frameworks, the best feature of EPSF is that it is typical in many ways and its related quantities are expressed in relatively compact forms. Another useful example is the Gibbs partition derived from the normalized inverse Gaussian process (Lijoi et al. 2005, 2007, 2008).

4 Subsampling from number partitions

Since the size index fluctuates with its total size n, to compare two or more random partitions with a different total size, some authors propose to reduce larger ones to the minimum by simple random sampling of individuals without replacement. Conceptually, a couple of ways of sampling is conceivable, and two of them are discussed in this section. The main concern is the mean number of parts in each situation.

In the first subsection, the basic sampling from a given number partition is discussed. In the second subsection, the sampling from a conditional random partitions on $\mathscr{P}_{\nu,\kappa}$ is discussed.

4.1 Sampling from a number partition

Regard a partition $z = (z_1, \ldots, z_\kappa) \in \mathscr{P}_{\nu,\kappa}$ as a set of κ urns with z_i balls, $\sum_{i=1}^{\kappa} z_i = \nu$. Take out one ball at random with equal probability and continue the sampling without replacement until n balls are obtained. Let W_i denote the number of balls taken from the ith urn, and $W = (W_1, \ldots, W_\kappa)$ follows the multivariate hypergeometric distribution MvHg$(n; z)$. Consider the size index of z, and partial sums of W_i along the size (value) of z_i:

$$Y = (Y_1, \ldots, Y_\nu),\ Y_k := \sum_{\{i; z_i = k\}} W_i,$$

$$\tau = (\tau_1, \tau_2, \ldots),\ \tau_k := \sum_{i=1}^{\kappa} \mathbb{I}[z_i = k], \quad \sum_{k=1}^{\nu} \tau_k = \kappa,\ \sum_{k=1}^{\nu} k\tau_k = \nu,$$

and Y follows MvHg$(n; \tau)$:

$$\mathbb{P}\{Y = y\} = \prod_{k=1}^{\nu} \binom{k\tau_k}{y_k} \Big/ \binom{\nu}{n}, \quad y = (y_1, , \ldots, y_\nu).$$

The conditional distributions of those W_is which are grouped to Y_k are

$$\mathbb{P}\{(W_i; z_i = k) = (w_i) | Y_k = y_k\} = \prod_{\{i; z_i = k\}} \binom{k}{w_i} \Big/ \binom{k\tau_k}{y_k}.$$

It is symmetric in $(w_i) := (w_i; z_i = k)$ and its order is irrelevant. Hence, the size index of $(W_i; z_i = k)|(Y_k = y_k)$ is introduced:

$$U_\ell := \sum_{i=1}^{\kappa} \mathbb{I}[z_i = k \ \&\ W_i = \ell],\ 0 \le \ell \le y_k, \ \text{ for } \ \tau_k > 0.$$

To avoid double index, we introduce the generic parameter (m, c, r) for (y_k, τ_k, k), and the *symmetric multivariate hypergeometric distributions* are defined by

$$\mathbb{P}\{U = u\} = c! \prod_{\ell=0}^{r} \frac{1}{u_\ell!}\binom{r}{\ell}^{u_\ell} \Big/ \binom{rc}{m},$$
$$u = (u_0, u_1, \ldots, u_r),\ \sum_{\ell=0}^{r} u_\ell = c,\ \sum_{\ell=0}^{r} \ell u_\ell = m,$$

and denoted by SymMvHg$(m; c, r)$. Note that U is a random partition of the number rc into nonnegative integers. Its alternative expression is the $2 \times c$ contingency table with such marginals that all c row-sums are equal to r and column-sums are m and $rc - m$.

The joint moments of U are

$$E\left(\prod_{\ell=0}^{r}(U_\ell)_{q_\ell}\right) = (c)_q \frac{(m)_Q(rc-m)_{rq-Q}}{(rc)_{rq}} \prod_{\ell=0}^{r}\binom{r}{\ell}^{q_\ell}$$
$$q_\ell \in \mathbb{N}_0,\ q := \sum_{\ell=0}^{r} q_\ell,\ Q := \sum_{\ell=0}^{r} \ell q_\ell. \tag{26}$$

To show it, note that

$$(c-q)! \prod_{\ell=0}^{r} \tfrac{1}{(u_\ell - q_\ell)!}\tbinom{r}{\ell}^{u_\ell - q_\ell} = \tbinom{r(c-q)}{m-Q},$$
$$\tbinom{r(c-q)}{m-Q}/\tbinom{rc}{m}. = \tfrac{(m)_Q(rc-m)_{rq-Q}}{(rc)_{rq}}.$$

This moment expression is a special case of Proposition 5, the binomial compound of the truncated binomial distributions.

The above discussions are summarized as below.

Proposition 6 *Given a partition $\tau \in \mathscr{P}_{\nu,\kappa}$, let the simple random sample of size n from τ be denoted by $S_{k\ell} = \sum_{i=1}^{\kappa} \mathbb{I}[W_i = \ell \ \&\ z_i = k]$, $0 \le \ell \le k$, namely the number of urns with ℓ balls taken from an urn with k balls, then*

$$\mathbb{P}\{(S_{k\ell}) = (s_{k\ell})\} = \prod_{k=1}^{\nu} \tau_k \prod_{\ell=0}^{k} \frac{1}{s_{k\ell}!}\binom{k}{\ell}^{s_{k\ell}} \binom{\nu}{n},$$
$$\sum_{\ell=0}^{k} s_{k\ell} = \tau_k, \quad \sum_{k=1}^{\nu}\sum_{\ell=0}^{k} \ell s_{k\ell} = n. \tag{27}$$

Proof Apply

$$\mathbb{P}\{(S_{k\ell}) = (s_{k\ell})\} = \sum_{y} \mathbb{P}\{(S_{k\ell}) = (s_{k\ell})|Y = y\}\mathbb{P}\{Y = y\}$$

to the discussions above the proposition. In $\sum_y$, y runs over a subset of $\mathscr{P}_n$. □

From Proposition 6, the p.m.f. or the moments of the statistics of $S_{k\ell}$ are calculated. For example, the marginal distribution of subsamples, in terms of the size index $S_{.\ell} := \sum_{k=\ell}^{\nu} S_{k\ell},\ 0 \le \ell \le n$, is

$$\mathbb{P}\{(S_{.\ell}) = (s_\ell)\} = \sum^{*} \tau_k! \prod_{\ell=0}^{k} \frac{1}{s_{k\ell}!}\binom{k}{\ell}^{s_{k\ell}} \Big/ \binom{\nu}{n},$$

where the summation $\overset{*}{\Sigma}$ runs over $s_{k\ell}$ satisfying $\sum_{k=1}^{\nu} s_{k\ell} = s_\ell$ and the restrictions in (27). The enumeration is complex.

The joint moments of $(S_{k\ell})$ are calculated by

$$E((S_{k\ell})_{q_{k\ell}}) = \sum_{y} E((S_{k\ell})_{q_{k\ell}}|Y = y)\mathbb{P}\{Y = y\},$$

where the conditional expectation is that of SymMvHg. Among them the following is the simplest and interesting.

Proposition 7 *The statistics* $K_n(\tau) := \sum_{k=1}^{\nu}\sum_{\ell=1}^{k} S_{k\ell}$ *of the subsample of* τ*, defined in Proposition 6, is the number of parts, or species, in the subsample.*

$$E(K_n(\tau)) = \sum_{k=1}^{\kappa} \tau_k \left(1 - \frac{(\nu - n)_k}{(\nu)_k}\right) = \kappa - \mu(\tau), \quad \mu(\tau) := \sum_{k=1}^{\kappa} \tau_k \frac{(\nu - n)_k}{(\nu)_k},$$

$$Var(K_n(\tau)) = \sum_{k=1}^{\kappa} (\tau_k)_2 \frac{(\nu - n)_{2k}}{(\nu)_{2k}} + 2 \sum_{1 \le j < k \le \nu} \tau_j \tau_k \frac{(\nu - n)_{j+k}}{(\nu)_{j+k}} - \mu(\tau)^2 + \mu(\tau).$$

The expectation shows how the number of parts $\kappa = \sum_{\ell=1}^{\nu} \tau_\ell$ *decreases with the sampling ratio* $\rho = n/\nu$*. Note that*

$$\frac{(\nu - n)_k}{(\nu)_k} = (1 - \rho)^k \exp\left(-\frac{k(k-1)}{2\nu}\frac{\rho}{1-\rho} + O(\nu^{-2})\right), \quad (\nu \to \infty).$$

Proof The conditional expectation of U_0 given $Y_k = m$ is obtained from (26) by putting $q_\ell = \mathbb{I}[\ell = 0],\ q = 1,\ Q = 0$, and Y_k has the marginal $\mathrm{Hg}(n; rc, \nu - rc),\ r = k, c = \tau_k$, of SymMvHg $(n; \tau)$. Hence

$$E(S_{k0}) = \left[\sum_{m=0}^{n} \frac{c}{(rc)_r}(rc-m)_r \binom{n}{m}\binom{\nu-n}{rc-m}\right] \Big/ \binom{\nu}{rc}$$
$$= \frac{c(\nu-n)_r}{(rc)_r}\binom{\nu-r}{rc-r}\Big/\binom{\nu}{rc} = c\frac{(\nu-n)_r}{(\nu)_r}.$$

What we need is $\kappa - \sum_{k=1}^{\kappa} E(S_{k0}) = \sum_{k=1}^{\kappa}(\tau_k - E(S_{k0}))$. The variance is similarly calculated. □

In the same way, it is shown that

$$E(S_{k\ell}) = \tau_k \binom{k}{\ell} \frac{(n)_\ell(\nu-n)_{k-\ell}}{(\nu)_k}, \quad 0 \le \ell \le k.$$

Check that $\sum_{\ell=0}^{k} E(S_{k\ell}) = \tau_k$.

The coefficients of τ_k in Proposition 7 are close to the limit $1-(1-\rho)^k$ even for smaller values of ν and n as shown in Fig. 1. In each of four frames, the differences between three curves, for $n = 16, 32, 64$, are invisible except for the jumps to the conventional value 1 outside the range $1 \le k \le n$. In the last frame, the true and approximate values at $k = n$ are as follows:

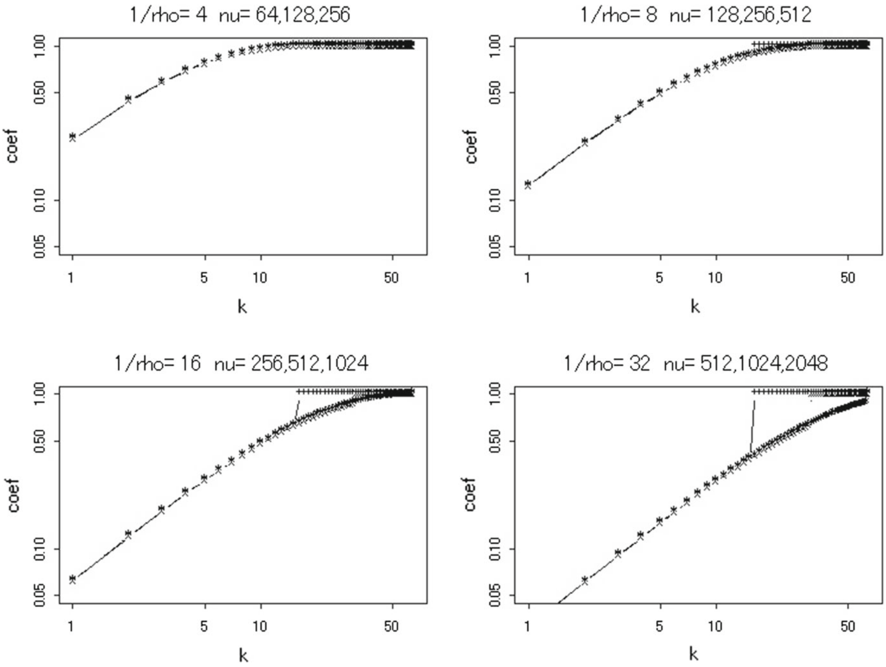

Fig. 1 The coefficient $(1-(\nu-n))/(\nu)_k$ *in* $E(K_n(\tau))$, versus $k = 1, 2, \ldots$ in log − log scale, for $n = 16, 32, 64$ in each frame with the sampling ratio $\rho = n/\nu = 1/4, 1/8.1/16, 1/32$

k	$n = 16$	$n = 32$	$n = 64$	$(1 - 1/32)^k$
16	0.493	0.401	0.399	0.398
32	1.0	0.644	0.641	0.638
64	1.0	1.0	0.873	0.869

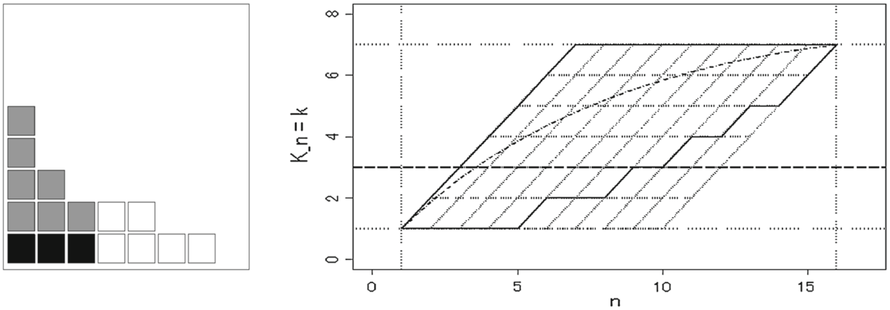

Fig. 2 *Left* Illustration of c of the example. *Right* Bounds of (n, K_n) paths. *Dotted line* Parallelograms show possible paths and a chain line curve shows $(n, E(K_n))$, calculated by Proposition 7

Numerical examples of Proposition 7 are shown in the following paragraph and in Sect. 5.

Bounds of K_n in subsampling To see how fast K_n will change in the subsampling of a given partition, we examine deterministic (or probability 1) bounds of $K_n, n = \nu-1, \nu-2, \ldots, 1$. In this paragraph, we discuss the unsampled balls, not the removed balls. Let $c = (c_1, \ldots, c_\kappa), c_1 \geq c_2 \geq \cdots \geq c_\kappa > 0, \sum_{i=1}^{\kappa} c_i = \nu$, be the DOS of a partition of ν, which is expressed by the vertical bar chart with columns of height c_j, or stacked c_j squares, at $j, \; 1 \leq j \leq \kappa$. Out of these ν squares, remove squares one by one, keeping the form of DOS: non-increasing height of columns.

Example (Fig. 2) Let $\nu = 16, \kappa = 7, c = (5, 3, 2, 2, 2, 1, 1)$ and consider, say n as $K_n = 3$. See the left part of Fig. 2. The largest n, or the earliest time in one-by-one sampling, is the case where the white squares are removed and both the black and the gray squares remain. The smallest n, or the latest time, is the case where the white and gray squares are removed and the black squares remain. Hence, $3 \leq n \leq 10$. If there are no white squares, $K_n = \kappa$, the upper bound remains κ while $n > \kappa$. The lower bound decreases by 1, from i to $i - 1$, if a column of original height c_i disappears.

Changing n the right part of Fig. 2 is obtained. The upper and the lower bounds n of $K_n = k, k$ specified, are the lower and the upper bounds of K_n, n specified, respectively. These facts are summarized as the following proposition.

Proposition 8 *In the subsample of size n of the partition data $c = (c_1, \ldots, c_\kappa)$, $\sum_{i=1}^{\kappa} c_i = \nu$, $n|(K_n = k)$ satisfies,*

$$\min(k, \kappa) \leq n|(K_n = k) \leq \sum_{i=1}^{k} c_i, \quad 1 \leq k \leq \kappa.$$

with the probability 1. Equivalently

$$L_c(n) \le K_n \le \min(n, \kappa), \quad 1 \le n \le \nu, \quad L_c(n) := k, \ \text{if} \ \sum_{i=1}^{k-1} c_i < n \le \sum_{i=1}^{k} c_i,$$

with the probability 1.

The limits are rather restrictive if ν is not large.

4.2 Sampling from random partitions

Sampling from the conditional random partitions $S|(S \in \mathscr{P}_{\nu,\kappa})$ In Sect. 3.1, the conditional distribution and its moments of the random sum models were obtained. In the present situation, Proposition 5 is applied as follows:

Proposition 9 *Assume the conditional random partition $S|(S \in \mathscr{P}_{\nu,\kappa})$ with the compounded exponential generating function $\breve{h}(w)$, and consider the simple random subsample of the size n from the partition. Then*

$$E(K_n|S \in \mathscr{P}_{\nu,\kappa}) = \sum_{k=1}^{\kappa} E(T_k|S \in \mathscr{P}_{\nu,\kappa}) \left(1 - \frac{(\nu - n)_k}{(\nu)_k}\right),$$

$$E(T_k|S \in \mathscr{P}_{\nu,\kappa}) = (\nu)_k \frac{\mathrm{B}_{\nu-k,\kappa-1}(\breve{h})}{\mathrm{B}_{\nu,\kappa}(\breve{h})} \frac{\breve{h}_k}{k!}.$$

For EPSF(θ, α), the summands are the zero-truncated (Engen's) extended negative binomial ($0 < \alpha < 1$), the log-series ($\alpha = 0$), or the negative hypergeometric ($\alpha < 0$) r.v.'s, and

$$\breve{h}_k = (1 - \alpha| - 1)_{k-1}, \quad \mathrm{B}_{\nu,\kappa}(\breve{h}) = S_{\nu,\kappa}(-1, -\alpha, 0).$$

This is true, not only for EPSF(θ, α), but also for all the consistent random partitions of the Gibbs form, as remarked in Sect. 3.2.

EPSF *and the consistent Gibbs case* In Sect. 3.2, the consistency of Gibbs partitions was introduced.

Suppose random partitions on $\mathscr{P}_n$ have the consistent Gibbs form (24). Then, its *unconditional* simple random samples of size m has the same Gibbs form on $\mathscr{P}_m$.

This notion answers the question raised in the beginning of this section; the subsampling for decreasing the sample size is allowed at least in the consistent Gibbs forms. Repeated subsamples can be used for better and easier inference. Note that the simple random sampling of one ball from random partitions is the choice from the sequenced balls under the consistency assumption.

For EPSF(θ, α), it is known (6) that

$$w(s; n, k) := \mathbb{P}\{S = s|(S \in \mathscr{P}_{\nu,\kappa}\} = \frac{1}{S_{n,k}(-1, -\alpha, 0)} \prod_{j=1}^{n} \frac{1}{s_j!} \left(\frac{(1 - \alpha| - 1)_{j-1}}{j!}\right)^{s_j}.$$

As a reverse of RsEPSF, we study the sampling from $w(s;\nu,\kappa)$. Again, note that $w(s;\nu,\kappa)$ is the conditional distribution for all the consistent Gibbs partitions.

Proposition 10 (The simple random sampling from $w(s;\nu,\kappa)$) *Let* $s=(s_1,\ldots,s_\nu)\in\mathscr{P}_{\nu,\kappa}$*, and take out one ball from s with the equal probability* $1/\nu$*. For s of* $w(s;\nu,\kappa)$*:*

(a) *If* $s_1>0$*, and one ball is chosen from* s_1 *balls in* s_1 *urns containing just one ball (singletons), the result is the random partition* $w(s;\nu-1,\kappa-1)$ *and this occurs with the probability*

$$\frac{S_{\nu-1,\kappa-1}(-1,-\alpha,0)}{S_{\nu,\kappa}(-1,-\alpha,0)}.$$

(b) *Otherwise, the result is the random partition* $w(s;\nu-1,\kappa)$ *and this occurs with the probability*

$$\frac{(\nu-1-\kappa\alpha)S_{\nu-1,\kappa}(-1,-\alpha,0)}{S_{\nu,\kappa}(-1,-\alpha,0)}.$$

Remark: This is a probabilistic interpretation of the recurrence

$$\begin{aligned}S_{\nu,\kappa}(-1,-\alpha,0)&=S_{\nu-1,\kappa-1}(-1,-\alpha,0)\\&\quad+(\nu-1-\kappa\alpha)S_{\nu-1,\kappa}(-1,-\alpha,0),\quad 1\le\kappa\le\nu.\end{aligned}$$

The result is a mixture of two disjoint random partitions $w(s;\nu-1,\kappa-1)$ and $w(s;\nu-1,\kappa)$, $1<k<\kappa$. The probabilities are equal to 1, if $k=\kappa$ in case (a), and if $k=1$ in case (b).

Proof First, note that

$$\{(s_1,s_2,\ldots,s_\nu);\,(s_1+1,\ldots,s_\nu)\in\mathscr{P}_{\nu,\kappa}\}=\mathscr{P}_{\nu-1,\kappa-1},$$

and that

$$\{(s_1,\ldots,s_\nu);\,(s_1,\ldots,s_{j-1}-1,s_j+1,\ldots,s_\nu)\in\mathscr{P}_{\nu,\kappa},\ j=2,\ldots,\nu\}=\mathscr{P}_{\nu-1,\kappa}.$$

Part a: Under the condition $S=s$, one of s_1+1 balls is chosen with the probability $(s_1+1)/\nu$, $s_1=0,1,\ldots$. Hence, this happens in $w((s_1+1,s_2,\ldots,s_\nu)$ with the probability

$$\begin{aligned}&\sum_{(s_1+1,s_2,\ldots,s_\nu)\in\mathscr{P}_{\nu,\kappa}}\frac{s_1+1}{\nu}w((s_1+1,s_2,\ldots,s_\nu);\nu,\kappa)\\&=\sum_{s\in\mathscr{P}_{\nu-1,\kappa-1}}\frac{(\nu-1)!}{S_{\nu,\kappa}(-1,-\alpha,0)}\prod_{j=1}^{\nu-1}\frac{1}{s_j!}\left(\frac{(1-\alpha|-1)_{j-1}}{j!}\right)^{s_j}\\&=\frac{S_{\nu-1,\kappa-1}(-1,-\alpha,0)}{S_{\nu,\kappa}(-1,-\alpha,0)}.\end{aligned}$$

Part b: In other cases, one of $j(s_j+1)$ balls is chosen and s_j+1 turns to s_j+1 while $s_{j-1}-1$ turns to s_{j-1}, $s_{j-1}=1,2,\ldots,s_j=0,1,\ldots$. This happens in $w(s_j^*;\nu,\kappa)$, $s_j^*=(s_1,\ldots,s_{j-1}-1,s_j+1,\ldots,s_\nu)$, with the probability

$$\sum_{j=2}^{\nu}\left(\sum_{s_j^*\in\mathscr{P}_{\nu,\kappa}}\frac{j(s_j+1)}{\nu}w(s_j^*;\nu,\kappa)\right)$$
$$=\sum_{j=2}^{\nu}s_{j-1}\frac{(1-\alpha|-1)_{j-1}}{(1-\alpha|-1)_{j-2}}\sum_{s\in\mathscr{P}_{\nu-1,\kappa}}\frac{(\nu-1)!}{S_{\nu,\kappa}(-1,-\alpha,0)}\prod_{j=1}^{\nu}\frac{1}{s_j!}\left(\frac{(1-\alpha|-1)_{j-1}}{j!}\right)^{s_j}$$
$$=(\nu-1-\kappa\alpha)\frac{S_{\nu-1,\kappa}(-1,-\alpha,0)}{S_{\nu,\kappa}(-1,-\alpha,0)}.$$

□

Reverse EPSF-K *process* Based on the random sampling of Proposition 10, a backward equation, corresponding to the forward Eq. (4) of the EPSF-K process, is calculated:

$$f_n(k)=(n-k\alpha)\frac{S_{n,k}}{S_{n+1,k}}f_{n+1}(k)+\frac{S_{n,k}}{S_{n+1,k+1}}f_{n+1}(k+1),\quad 1\le k\le n,\qquad(28)$$

where $S_{n,k}=S_{n,k}(-1,-\alpha,0)$ for simplicity. For the proof, just recall $f_n(k)=(\theta|-\alpha)_k S_{n,k}/(\theta|-1)_n$.

Starting from the state $K_\nu=\kappa$, the conditional downward random walks are determined by modifying the boundary restrictions in the above (28). Note that the possible states are the parallelogram with the corners (ν,κ), $(\nu-\kappa+1,1)$, (κ,κ), $(1,1)$, or

$$\mathscr{P}^D(\nu,\kappa):=\bigcup_{k=1}^{n}\mathscr{P}_n^D(\nu,\kappa),$$
$$\mathscr{P}_n^D(\nu,\kappa):=\{(n,k);\max(1,n-(\nu-\kappa))\le k\le\min(n,\kappa)\},\ 1\le n\le\nu.$$

Proposition 11 (Reverse EPSF-K starting from $K_\nu=\kappa$) *Assume that* EPSF-K *process is at the state $K_\nu=\kappa$, that is $S\in\mathscr{P}_{\nu,\kappa}$ in* EPSF *having $w(s;\nu,\kappa)$. If one ball is deleted at random, one by one with the equal probability $1/\nu, 1/(\nu-1),\ldots$, the number of occupied urns K_n, $=\nu,\nu-1,\ldots$ decreases downward within the region $\mathscr{P}^D(\nu,\kappa)$, along the following distributions, normalized for each n.*

$$g_n(\kappa-\nu+n)=\frac{S_{n,\kappa-\nu+n}}{S_{n+1,\kappa-\nu+n+1}}g_{n+1}(\kappa-\nu+n+1)),\quad \nu-\kappa<n\le\nu,$$
$$g_n(1)=\frac{n-n\alpha}{n-\alpha}g_{n+1}(1)+\frac{S_{n,1}}{S_{n+1,2}}g_{n+1}(2),\quad 1\le n\le\nu-\kappa,$$
$$g_n(k)=(n-k\alpha)\frac{S_{n,k}}{S_{n+1,k}}g_{n+1}(k)+\frac{S_{n,k}}{S_{n+1,k+1}}g_{n+1}(k+1),$$

$$\begin{aligned}
&1 < k < \min(\kappa, n),\\
g_n(\kappa) &= (n - \kappa\alpha)\frac{S_{n,\kappa}}{S_{n+1,\kappa}} g_{n+1}(\kappa), \quad \kappa \le n \le \nu,\\
g_n(n) &= \frac{2}{n+1} g_{n+1}(n) + g_{n+1}(n+1), \quad n < \kappa.\\
g_\nu(\kappa) &= g_1(1) = 1, \qquad S_{n,k} = S_{n,k}(-1, -\alpha, 0).
\end{aligned}$$

Remarks

(a) Actually, the equation $g_n(k)$ in the middle of the lines, or (28), covers all the others, because $S_{n,n}(a, b, 0) = 1, \quad n = 1, 2, \ldots, S_{n+1,1}(a, b, 0) = (b - na)S_{n,1}(a, b, 0)$, and $S_{n,n-1}(a, b, 0) = 2(b - a)/(n(n - 1))$.
(b) Note that the process depends only on α, in contrast to EPSF-K.
(c) The means of $g_n(k), k \in \mathscr{P}_n^D(\nu, \kappa)$, correspond to those in Proposition 9.

Simple cases $\nu, \kappa = \nu - 1$

$$\begin{aligned}
g_\nu(\nu - 1) &= g_1(1) = 1,\\
g_{\nu-1}(\nu - 2) &= \frac{S_{\nu-1,\nu-2}}{S_{\nu,\nu-1}} = \frac{\nu - 2}{\nu}, \quad g_{\nu-1}(\nu - 1) = 1 - g_{\nu-1}(\nu - 2) = \frac{2}{\nu},\\
g_n(n - 1) &= \frac{S_{n,n-1}}{S_{n,n;;1}} g_{n+1}(n) = \frac{n-1}{n+1} g_{n+1}(n)\\
&= \frac{(n-1)(n-2)(n-3)}{(n+1)n(n-1)} \cdots \frac{\nu - 2}{\nu} \cdot 1 = \frac{n(n-1)}{\nu(\nu - 1)}, \quad 2 \le n \le \nu,\\
g_n(n) &= 1 - g_n(n - 1) = 1 - \frac{n(n-1)}{\nu(\nu - 1)}.
\end{aligned}$$

The last expression is directly derived from Proposition 11:

$$\begin{aligned}
g_n(n) &= n(1 - \alpha)\frac{S_{n,n}}{S_{n,n+1}} g_{n+1}(n) + \frac{S_{n,n}}{S_{n+1,n+1}} g_{n+1}(n + 1)\\
&= \frac{2}{n+1} g_{n+1}(n) + g_{n+1}(n + 1) = \frac{2n}{(\nu - 1)\nu} + 1 - \frac{n(n-1)}{\nu(\nu - 1)}\\
&= 1 - \frac{(n+1)n}{\nu(\nu - 1)}.
\end{aligned}$$

In this special case, the probabilities are independent of α. If $\nu = 5, \kappa = 4$,

$$\begin{aligned}
&g_4(3) = 0.6, \quad g_4(4) = 0.4 : \quad g_3(2) = 0.3, \quad g_3(3) = 0.7;\\
&g_2(1) = 0.1, \quad g_2(2) = 0.9.
\end{aligned}$$

The cases $\nu = 5, \kappa = 3$ & 2 are tabulated as follows:

n	$g_n(k)\times 5(7-5\alpha)$			n	$g_n(k)\times 5(5-3\alpha)$	
	$g_n(1)$	$g_n(2)$	$g_n(3)$		$g_n(1)$	$g_n(2)$
4		$11-7\alpha$	$2(12-9\alpha)$	4	$3-\alpha$	$2(11-7\alpha)$
3	$2-\alpha$	$3(7-5\alpha)$	$3(4-3\alpha)$	3	$7-3\alpha$	$6(3-2\alpha)$
2	$3(3-2\alpha)$	$26-19\alpha$		2	$13-7\alpha$	$4(3-2\alpha)$

It is conjectured that the subsamples of Proposition 11 of size n, under the condition (ν, κ) where κ is close to $E(K_\nu)$ of EPSF-K, have the mean close to $E(K_n)$. An application of Proposition 11 is shown in the following Sect. 5. In numerical computation of $S_{n,k}$ for large n, a way to avoid overflow is to use EPSF-K$(0, \alpha)$, (4) (5):

$$S_{n,k}(-1, -\alpha, 0) = \alpha^{-k}(n-1)!f_n(k)/(k-1)!$$

Once $\{g_n(k);\ \ (n,k) \in \mathscr{P}^D(\nu,\kappa)\}$ are obtained, the transition probabilities of the upward random walks (n, K_n), $n = 1\ldots,$ of EPSF-K(θ, α), under the condition $K_\nu = \kappa$, are easily determined. Let $p(k|k)$ and $p(k+1|k)$ denote the move of probabilities from (n,k) to $(n+1,k)$ and $(n+1,k+1)$, that is $p(k|k)+p(k+1|k) = g_n(k)$ and the transition probabilities are $p(k|k)/g_n(k)$ and $p(k+1|k)/g_n(k) = 1 - p(k|k)/g_n(k)$. From the boundary $p(1|1) = g_{n+1}(1)$ and $p(2|1) = g_n(1) - p(1|1))$, the others are calculated by $p(k|k) = g_{n+1}(k) - p(k|k-1)$ and $p(k+1|k) = g_n(k) - p(k|k)$, $k = 2,\ldots$ Hence, the upward transition probabilities are independent of θ. The upward random walks have the same paths (chains) and the same moving probabilities as the downward random walks.

Further, consider the downward random walks starting from (ν, κ) and ending at any point $(\mu, \lambda) \in \mathscr{P}^D(\nu, \kappa)$, or $1 < \mu < \nu$, $\max(1, \mu - (\nu - \kappa)) \le \lambda \le \min(\mu.\kappa)$, the conditional downward random walks, with both ends fixed, have the same features as those on $\mathscr{P}^D(\nu, \kappa)$. This fact is discussed in Gnedin and Pitman (2006), Section 4.

Subsampling from RsEPSF For the forward Eq. (4) of EPSF-K, the backward equation is given by (28). Similarly, for the forward Eq. (10) of $K_n(m,k)$ or RsEPSF-K, the backward equation is as follows:

$$f_n(\ell) = (m+n-(k+\ell)\alpha)\frac{S_{n,\ell}}{S_{n+1,\ell}}f_{n+1}(\ell) + \frac{S_{n,\ell}}{S_{n+1,\ell+1}}f_{n+1}(\ell+1), \quad 0 \le \ell \le n, \tag{29}$$

where $S_{n,k} = S_{n,k}(-1, -\alpha, m - k\alpha)$ for simplicity. Proposition 11 is similarly extended.

Next, consider the Markov process in the finer states $(n; \ell, s_0)$. From the forward Eq. (19) in Sect. 2.2, the backward equation, or simple random deletion of one ball, is as follows. Here, $S_{n,k} = S_{n,k}(-1, -\alpha, 0)$ and $n^* + s_0 = n$:

Table 1 Size index $s = (s_1, s_2, \dots)$ of trawl byproducts (Heales et al. 2003a)

j	1	2	3	4	5	6	7	8	9	10	11	13
s_j	22	10	6	6	3	2	2	7	3	2	2	4
j	14	15	16	17	18	19	20	29	31	36	38	40
s_j	1	3	1	2	2	2	1	1	1	1	1	1
j	47	49	55	61	73	78	90	112	129	138	187	189
s_j	2	1	1	1	1	1	1	1	1	1	1	1
j	201	236	251	252	293	328	353	363	531	539	558	722
s_j	1	1	1	1	1	1	1	1	1	1	1	1
j	731	890	1380	1926	2123							
s_j	1	1	1	1	1							

$(\sum_j s_j = 116, \sum_j js_j = 13611)$

$$f_n(\ell, s_0) = \frac{n^*+1}{n+1}\left((n^* - \ell\alpha)\frac{S_{n^*,\ell}}{S_{n^*+1,\ell}} f_{n+1}(\ell, s_0) + \frac{S_{n^*,\ell}}{S_{n^*+1,\ell+1}} f_{n+1}(\ell+1, s_0)\right) + \frac{s_0+1+1}{n+1}\frac{S_{n^*,\ell}}{S_{n^*+1,\ell}} f_{n+1}(\ell, s_0+1) \quad 0 \le \ell \le n. \tag{30}$$

5 Application to trawl data

Trawl data In tropical northern Australia coast, the Northern Pawn Fishery is supporting ecologically sustainable development, and controlling trawl bycatch. Studying the effect of the control, important issues were techniques of homogeneous subsampling and random fluctuation of subsamples. To answer the questions, large-scale works were carried out and reported by Heales et al. (2000, 2003a,b). In this paper, only the fluctuation of species abundance is reexamined, and one of the datasets in Heales et al. (2003a) is analyzed. The datasets are large and very unique in that all bycatches are sequentially subsampled and species of all individuals in each subsample are enumerated. Each subsample has almost the same weight. One of the three datasets in Heales et al. (2003a), Catch no. 2 with the largest number of individuals, is used to support the proposed model. The size index, $s = (s_1, s_2, \dots)$, of the whole sample is listed in Table 1. The dataset was analyzed by Shimadzu and Darnell (2013), which motivated the author to carry out this study.

The sample consists of 26 boxes of about 10 kg of fish and invertebrates. The number of individuals n and the accumulated number of species K_n accumulated up to the box number 'bxn', bxn = 1, ..., 26, are listed in Table 2. The last column, bxn = 26, is the whole sample used in Table 1. Size index, as in Table 1, was prepared also for bxn = 8 and 16. Summary of these subsamples and maximum likelihood estimates $(\hat{\theta}, \hat{\alpha})$ is provided in Table 3.

Fitting EPSF In Fig. 3, the expectation $E(K_n)$ of EPSF-K $(\hat{\theta}, \hat{\alpha})$, $1 \le n \le 20000$, is plotted by a solid line, with the curves $E(K_n) \pm 2SD(K_n)$ by broken lines. $E(K_n)$ and $SD(K_n)$ are computed by MLE $(\hat{\theta}, \hat{\alpha})$ of the whole sample in Table 3. In Fig. 3, observed (n, K_n) of Table 2 are shown by symbol 'X'. They are so close to $(n, E(K_n))$.

Table 2 Subsampling process (Heales et al. 2003a)

bxn	1	2	3	4	5	6	7	**8**	9
K_n	49	64	68	74	81	84	87	89	90
n	607	1341	1948	2499	3039	3700	4183	4723	5229
bxn	10	11	12	13	14	15	**16**	17	18
K_n	92	94	95	98	102	107	107	109	112
n	5785	6207	6558	6994	7476	7883	8366	8918	9601
bxn	19	20	21	22	23	24	25	**26**	
K_n	113	114	115	115	115	115	116	116	
n	10175	10804	11405	11885	12377	12910	13258	13611	

Table 3 Estimated parameter values

No. boxes	K_n	n	$\hat{\theta}$	$\hat{\alpha}$
8	89	4723	7.48537	0.172137
16	107	8366	8.31265	0.163111
26	116	13611	10.05340	0.120449

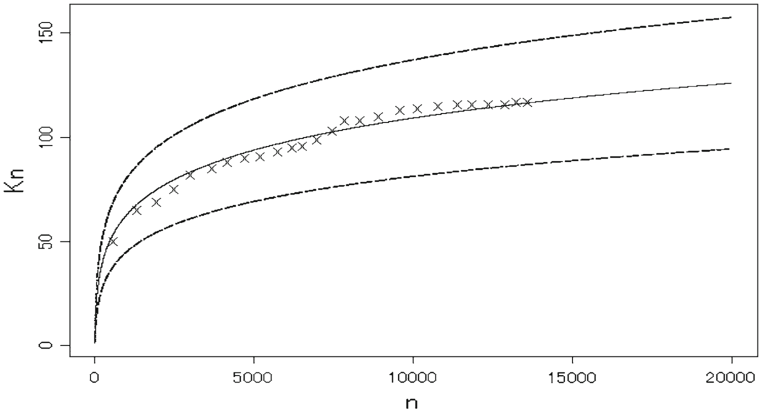

Fig. 3 (n, K_n) plots and $(n, E(K_n))$, $(n, E(K_n) \pm 2SD(K_n))$ curves. The parameter of $E(K_n)$ is an estimate of the last column of Table 3

In the middle of the train of 'X', there is a small jump. Heales et al. (2003a) examined details of data, and found that some rare species are sampled successively at that stage. However, compared with the possible fluctuation of EPSF-K, the jump is relatively small, and the subsampling process appears random and homogeneous. Figure 4 and Table 4 show the probability of the intervals separated by mean and standard deviation, to evaluate the performance of the two-sigma interval of Fig. 3.

Prediction In Fig. 5, $E(K_\ell(m, k))$, $K_\ell(m, k) = K_{m+\ell}|(K_m = k)$ and $E(K_\ell(m, k)) \pm 2SD$ are plotted, in solid and broken lines, respectively. Conditional expectation is very close to the unconditional expectation. If the conditioning value shifts, the conditional

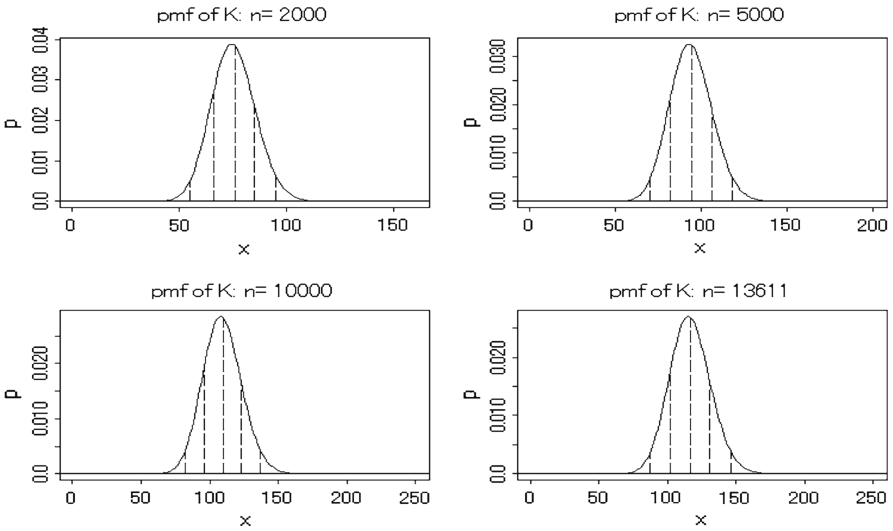

Fig. 4 P.d.f. of EPSF-K(θ, α), using estimates ($\hat{\theta}, \hat{\alpha}$) of the last column of Table 3. *Vertical broken lines* show $E(K_n)$, $E(K_n) \pm SD(K_n)$, $E(K_n) \pm 2SD(K_n)$

Table 4 Cumulative distribution functions of EPSF-K of the estimated parameter value

n	$x =$	$\lceil \mu - 2\sigma \rceil$	$\lceil \mu - \sigma \rceil$	$\lceil \mu \rceil$	$\lfloor \mu + \sigma \rfloor$	$\lfloor \mu + 2\sigma \rfloor$
5000	x	70	82	94	106	118
	$F(x)$	0.02367	0.17957	0.53282	0.84767	0.97282
10000	x	82	96	110	123	137
	$F(x)$	0.02259	0.18218	0.54688	0.84539	0.97364
15000	x	89	104	119	133	148
	$F(x)$	0.02005	0.16960	0.52775	0.83476	0.97081

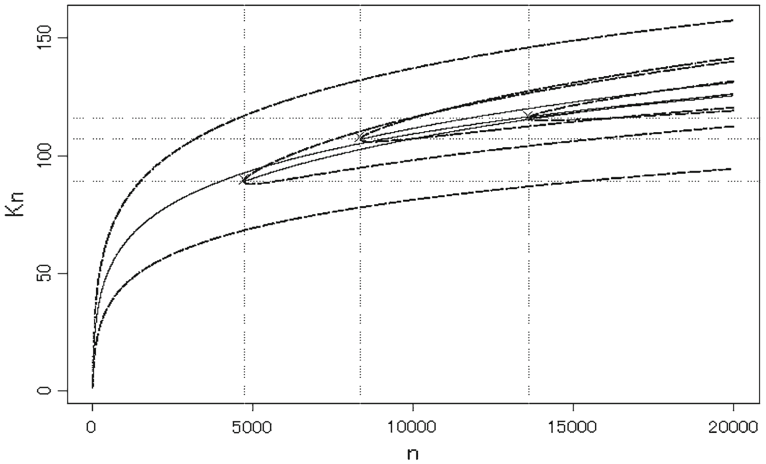

Fig. 5 Prediction of the number of species from the whole and subsamples, using estimates ($\hat{\theta}, \hat{\alpha}$) of Table 3

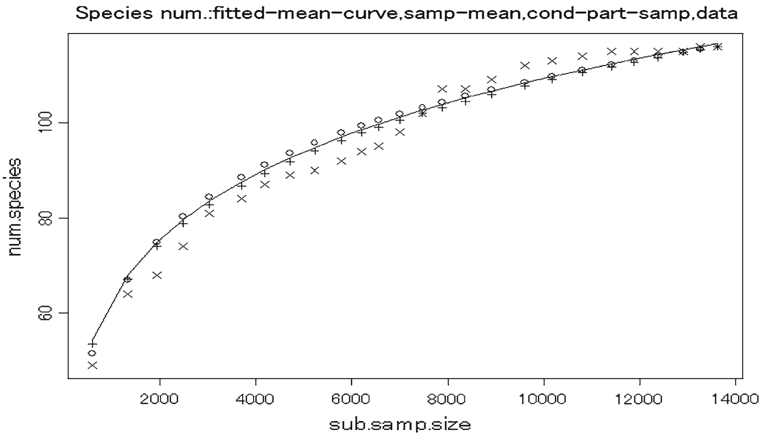

Fig. 6 Expected species numbers (*open circle*) of random samples from the data partition of Table 1, expected species numbers (*plus symbol*) of subsamples from the conditional random partitions of EPSF$(\cdot, \alpha)$ on $\mathscr{P}_{13611,116}$, with the observed species numbers (*cross symbol*) of Table 2 (*cross symbol*) and the fitted $(n, E(K_n))$ curve

expectation curve shifts almost in parallel. Further, smaller proportion of subsamples predicts the exact behavior of the larger proportion.

Subsamples Since the primary purpose of the trawl survey was to check the effects of subsampling on the number of species, it is instructive to apply Propositions 7 and 11 to this dataset. Figure 6 illustrates

($\circ$) The expected species number of random samples from the data partition of Table 1. See Proposition 7.

($+$) The expected species number of subsamples from the conditional random partitions of EPSF$(\cdot, \alpha)$ on $\mathscr{P}_{13611,116}$. See Proposition 11. The estimate $\hat{\alpha}$ in Table 3 for $n = 13611$ is used.

($\times$) The observed species numbers of Table 2.

Curve $(n, E(K_n))$ The curve of the fitted EPSF distribution and the observed number of species in subsamples.

The curve and ($\times$) are the same as Fig. 3. The subsample expectations, of both data partition and conditional random partitions, are surprisingly close to the fitted EPSF expectation curve.

Figure 7 illustrates the p.m.f.s of the subsamples from the conditional random partitions of EPSF$(\cdot, \alpha)$ on $\mathscr{P}_{13611,116}$. In the interval (1,116), the p.d.f curves concentrate rapidly to the neighborhood of 1 and are not exactly shown in this figure.

6 Final discussions

- This paper shows that EPSF is a useful candidate for modeling the species abundance data, and that G3SN is a useful tool for dealing with EPSF. In particular, the EPSF model is attractive for marine ecological surveys, where the subsampling is inevitable.

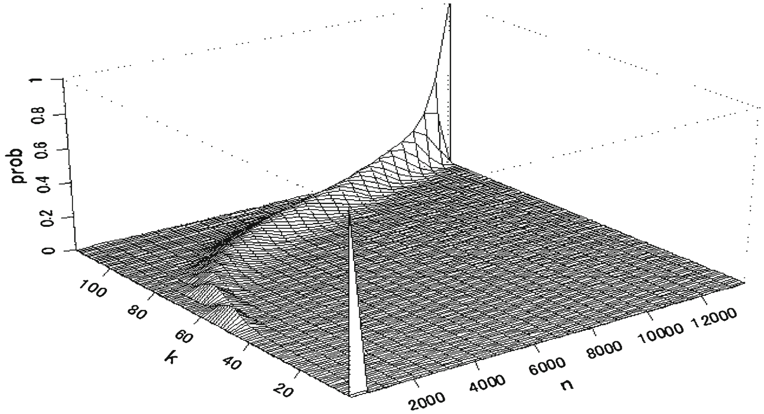

Fig. 7 The p.m.f.s of K_n, subsamples from the conditional random partitions of EPSF$(\cdot, \alpha)$ on $\mathscr{P}_{13611,116}$

- In modeling ecological surveys, one will assume that the parameter (θ, α) depends on covariates reflecting habitats and catching techniques, and that the estimated dependence will reveal insight into the real world. Species abundance and biological diversity continue to be a challenging problem. See, e.g., Hubbell (2001) and Guisan and Zimmermann (2000).
- On the other hand, for looking at *species abundance distributions* as random partitions, conceptual unification with other models will be required. See, e.g., McGill et al. (2007) for an extensive survey of models of partition data.
- The subsamples of a data partition are another expression of the partition. Hence, it is expected that subsamples will be used to select candidate random partition models.

Appendix: Generalized Stirling numbers

For the use in Sect. 2, recall the 3-parameter generalized Stirling numbers (G3SN) introduced by Hsu and Shiue (1998). It is a triangular sequence $S_{n,k}$, $0 \leq k \leq n$, of polynomials of order $n-k$ in a, b and c with integer coefficients, defined by the following equivalent conditions.

A (by a triangular generating function, specified by a pair of exponential generating functions)

$$S_{n,k}(a,b,c) := \left[\frac{t^n u^k}{n!}\right] \psi(t)\exp(u\phi(t)),$$
$$(\psi, \phi) := \left((1+at)^{c/a}, \frac{1}{b}((1+at)^{b/a} - 1)\right),$$

B (by a polynomial identity in t)

$$(t+c|a)_n \equiv \sum_{k=0}^{n} S_{n,k}(a,b,c)(t|b)_k,$$

C (by a recurrence formula)

$$S_{n+1,k}(a,b,c)=(kb-na+c)S_{n,k}(a,b,c)+S_{n,k-1}(a,b,c), \quad S_{0,0}(a,b,c)=1.$$

D (by an expansion, Corcino 2001)

$$S_{n,k}(a,b,c)=\frac{1}{b^k k!}\sum_{j=0}^{k}\binom{k}{j}(-1)^{j-k}(jb+c|a)_n.$$

This is a generalization of $S_{n,k}(1,0,0)=\left[\begin{matrix}n\\k\end{matrix}\right](-1)^{n-k}$, Stirling number of the first kind; $S_{n,k}(0,1,0)=\left\{\begin{matrix}n\\k\end{matrix}\right\}$, Stirling number of the second kind; and $S_{n,k}(0,0,1)=\binom{n}{k}$, binomial coefficients. Its basic properties are

$$\begin{aligned}
S_{n,k}(a,b,0) &= \mathbb{I}[n=k], \text{ if } a=b.\\
S_{n,k}(sa,sb,sc) &= s^{n-k}S_{n,k}(a,b,c), \ \forall s,\\
\sum_{\ell=k}^{n} S_{n,\ell}(a,s,c_1)S_{\ell,k}(s,b,c_2) &= S_{n,k}(a,b,c_1+c_2), \ \forall s.\\
S_{n,k}(a,b,c) &= \sum_{n\ge\ell\ge k}\binom{n}{\ell}S_{\ell,k}(a,b,0)(c|a)_{n-\ell}. \qquad (31)
\end{aligned}$$

In (8) in Proposition 1 , put $n=1$, $\ell=k$ to obtain Definition C. That is, (8) is another definition of G3SN. See, Hsu and Shiue (1998), Corcino (2001) and Wang and Wang (2008) for the details of G3SN.

Relationship with the Bell polynomials The special case $c=0$ of G3SN is related to the partial exponential Bell polynomials (Sect. 3.1) as follows:

$$\begin{aligned}
S_{n,m}(a,b,0) &= \mathrm{B}_{n,m}((b-a|a)_{k-1}) = b^{-m}\mathrm{B}_{n,m}((b|a)_k)\\
&= \frac{a^n}{b^m}\mathrm{B}_{n,m}\left(\left(\frac{b}{a}\right)_k\right) = a^{n-m}\mathrm{B}_{n,m}\left(\left(\frac{b}{a}-1\right)_k\right).
\end{aligned}$$

Noncentral generalized factorial coefficients Noncentral generalized factorial coefficients $C(\cdot)$ are defined by the polynomial identity in t;

$$(st+r)_n=\sum_{k=0}^{n}C(n,k;s,r)(t)_k.$$

See Charalambides and Singh (1988) or Charalambides (2005). The case $r=0$ is called generalized factorial coefficients. In terms of G3SN,

$$C(n,m;s,r)=s^m S_{n,m}(1,s,r)=s^n S_{n,m}(1/s,1,r/s), \quad 0\le m\le n.$$

For generalized factorial coefficients, in terms of the exponential partial Bell polynomials,

$$C(n, m; s, 0) = B_{n,m}(((s)_k)) = s^m B_{n,m}(((s-1)_{k-1})).$$

Lijoi et al. (2007, 2008) used the notation

$$(st + r| - 1)_n = \sum_{k=0}^{n} \mathscr{C}(n, k; s, r)(t| - 1)_k.$$

Hence,

$$\mathscr{C}(n, m; s, r) = s^m S_{n,m}(-1, -s, -r) = s^n S_{n,m}(-1/s, -1, -r/s).$$

In some context, G3SN variables are redundant, but G3SN is a typical Riordan array, e.g., Wang and Wang (2008), and known to relate to various popular polynomials.

Acknowledgments The author thanks the referees for useful comments improving the paper. The data in Sect. 5 were made available by courtesy of Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia. The author thanks the scientists and crews who participated in the 1997 RV Southern Surveyor voyages and CSIRO Marine and Atmospheric Research. He thanks Dr. Ross Darnell, CSIRO Mathematics, and Dr. Hideyasu Shimadzu, Geoscience Australia, for their guide to marine science and the subsampling problem. This work was supported by Grant-in-Aid for Scientific Research awarded by the Japan Society for the Promotion of Science.

References

Andrews, G.W., Eriksson, K. (2004). *Integer partitions*. Cambridge University Press, UK. (Japanese Translation by F. Sato, 2006, Tokyo: Sugaku Shobo.).

Carlton, M. A. (1999). *Applications of the two-parameter Poisson-Dirichlet distributions*. Ph.D. dissertation. Los Angeles: Department of Statistics, University of California.

Charalambides, C. A. (2002). *Enumerative combinatorics*. Hoboken, NJ: Wiley.

Charalambides, C. A. (2005). *Combinatorial methods in discrete distributions*. BocaRaton, FL: Chapman & Halls/CRC.

Charalambides, C. A. (2007). Distributions of random partitions and their applications. *Methodology and Computing in Applied Probability*, *9*, 163–193.

Charalambides, C. A., Singh, J. (1988). A review of the Stirling numbers, their generalizations and statistical applications. *Communication in Statistics-Theory and Methods*, *17*, 2533–2595.

Comtet, L. (1974). *Advanced combinatorics: the art of finite and infinite expansions*. Dordrecht, Netherlands: Reidel.

Corcino, R. B. (2001). Some theorems on generalized Stirling numbers. *Ars Combinatoria*, *60*, 273–286.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, *1*, 209–230.

Gnedin, A., Pitman, J. (2006). Exchangeable Gibbs partitions and Stirling triangles. *Mathematical Sciences*, *138*, 5674–5685. (original Russian version: Zapiski Nauchnnykh Seminarov ROMI, 325, 2005, 83–102.).

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, *40*, 237–264.

Guisan, A., Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, *135*, 147–186.

Heales, D. S., Brewer, D. T., Wang, Y.-G. (2000). Subsampling multi-species trawl catches from tropical northern Australia: Does it matter which part of the catch is sampled? *Fisheries Research*, *48*, 117–126.

Heales, D. S., Brewer, D. T., Jones, P. N. (2003a). Subsampling trawl catches from vessels using seawater hoppers: Are catch composition estimates biased? *Fisheries Research*, *63*, 113–120.

Heales, D. S., Brewer, D. T., Wang, Y.-G., Jones, P. N. (2003b). Does the size of subsamples taken from multispecies trawl catches affect estimates of catch composition and abundances? *Fishery Bulletin*, *101*, 790–799.

Hoshino, N. (2001). Applying Pitman's sampling formula to microdata disclosure risk assessment. *Journal of Official Statistics*, *17*, 499–520.

Hoshino, N. (2012). Random partitioning over a sparse contingency table. *Annals of the Institute of Statistical Mathematics*, *64*, 457–474.

Hsu, L. C., Shiue, P. J.-S. (1998). A unified approach to generalized Stirling numbers. *Advances in Applied Mathematics*, *20*, 366–384.

Hubbell, S. P. (2001). *The unified neutral theory of biodiversity and biogeography*. NJ: Princeton University Press.

Johnson, N. L., Kemp, A. W., Kotz, S. (2005). *Univariate Discrete Distributions* (3rd ed.). New York, NY: Wiley.

Kerov, S. V. (2006). Coherent random allocations, and the Ewns-Pitman formula. *Mathematical Sciences*, *135–3*, 5699–5710.

Lijoi, A., Mena, R. H., Prünster, I. (2005). Hierarchical mixture modeling with normalized inverse Gaussian priors, *Journal of the American Statistical Association*, *100*, 1278–1291.

Lijoi, A., Mena, R. H., Prünster, I. (2007). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika*, *94–4*, 769–786.

Lijoi, A., Prünster, I., Walker, S. G. (2008). Bayesian nonparametric estimators derived from conditional Gibbs structures. *The Annals of Applied Probability*, *18–4*, 1519–1547.

McGill, B.J., et al. (2007). Species abundance distributions: moving beyond single prediction theories to integration within an ecological frame work. *Ecology Letters*, *10*, 995–1015. (17 coauthors are abbreviated).

Pitman, J. (2006). *Combinatorial Stochastic Processes*. Lecture Notes in Mathematics, Vol. 1875. New York, NY: Springer.

Shimadzu, H., Darnell, R. (2013). Quantifying the effect of sub-sampling on species abundance distributions, (submitted).

Sibuya, M. (1993). A random clustering process. *Annals of the Institute of Statistical Mathematics*, *45*, 459–465.

Sibuya, M., Nishimura, K. (1997). Prediction of record-breakings. *Statistica Sinica*, *7*, 893–906.

Sibuya, M., Yamato, H. (2001). Pitman's model of random partitions. *RIMS Kokyuroku*. Research Institute for Mathematical Science, Kyoto University, *1240*, 64–73. (A revised version: *International Conference on Advances in Statistical Inferential Methods: Theory and Applications, Proceedings*, June 9–12, 2003, Kazakhstan Inst. Manag. Econ. Strat. Res. (KIMEP), Almaty, pp. 219–231).

van Ark, H., Meiswinkel, R. (1992). Subsampling of large light trap catches of culicoides (diptera: ceratopogonidae). *Ondelstepoort Journal of Veterinary Research*, *59*, 183–189.

Wang, W., Wang, T. (2008). Generalized Riordan arrays. *Discrete Mathematics*, *308*, 6466–6500.

Yamato, H., Sibuya, M. (2003a). Moments of some statistics of Pitman Sampling Formula. *Bulletin of Informatics and Cybernetics, Fukuoka*, *32*, 1–10.

Yamato, H., Sibuya, M. (2003b). Some topics on Pitman's random partition. *Proceedings of the Institute of Statistical Mathematics*, *51*, 351–372. (in Japanese).