doi: 10.1016/S0167-7152(01)00119-5

Statistics & Probability Letters 55 (2001) 19-27

ORDERED SAMPLE FROM TWO-PARAMETER GEM DISTRIBUTION

Hajime Yamato[a,*], Masaaki Sibuya[b] and Toshifumi Nomachi[c]

[a] *Dept. of Math. and Comp. Sci., Kagoshima Univ., Kagoshima 890-0065, Japan*

[b] *Takachiho University, Suginami-ku, Tokyo 168-8508, Japan*

[c] *Grad. School of Sci. and Engin., Kagoshima Univ., Kagoshima 890-0065, Japan: and Miyakonojo Col. of Tech., Miyazaki 885-8567, Japan*

Abstract. Pitman's two-parameter GEM distribution is characterized by the product moments. Based on that, sampling distributions of some statistics of a random sample from the two-parameter GEM distribution is derived. The main statistics are the frequencies of categories in a sample in the order of appearance, the intervals between new categories and number of distinct categories.

Key words and phrases: C-number, Donnelly-Tavaré-Griffiths formula, GEM distribution, Generalized Stirling numbers, Pitman sampling formula, Residual allocation model.

1 Introduction and GEM distribution

Let $(W_j)_{j=1}^{\infty}$ be a sequence of independent random variables, with W_j following the beta distribution $\mathsf{Be}\,(1-\alpha, \theta+j\alpha)$, $j \in \mathcal{N} = \{1, 2, \dots\}$, $(0 \le \alpha < 1,\ \theta > -\alpha)$. Let the standard simplex of the infinite dimension be denoted by

$$\Delta = \left\{ (p_j)_{j=1}^{\infty};\ p_j \ge 0,\ j \in \mathcal{N},\ \sum_{j\in\mathcal{N}} p_j = 1 \right\}.$$

A random vector $\boldsymbol{Z}= (Z_j)_{j=1}^{\infty}$ on Δ is a transform of $(W_j)_{j=1}^{\infty}$ by the residual allocation, Patil and Taillie (1977) and Pitman (1996a):

$$Z_1 = W_1, \quad \text{and} \quad Z_j = (1-W_1)\cdots(1-W_{j-1})\,W_j, \quad j = 2, 3, \dots \tag{1}$$

Note that $Z_1 + \cdots + Z_j + (1-W_1)\cdots(1-W_j) = 1$. The distribution of $\boldsymbol{Z}$, in the case $\alpha = 0$, is known as the GEM (Generalized Engen-McClosky) distribution. See, e.g. Johnson et al. (1997) or Pitman and Yor (1997). Hence, we call the distribution of $\boldsymbol{Z}$, defined by (1), the *two-parameter GEM distribution*, and express it as $\mathsf{GEM}\,(\theta, \alpha)$.

Regard $\boldsymbol{Z}$ of $\mathsf{GEM}\,(\theta, \alpha)$ as a sequence of random probabilities of an infinite number of categories indexed by $\mathcal{N}$, and let $\boldsymbol{X}= (X_1, \dots, X_n)$ be a random sample from $\boldsymbol{Z}$, that is, an i.i.d. sequence on $\mathcal{N}^n$, with the random probabilities:

$$P\left(\boldsymbol{X} = (x_1, \dots, x_n) \mid \boldsymbol{Z} = (z_1, z_2, \dots)\right) = \prod_{j=1}^{n} z_{x_j}, \quad x_j \in \mathcal{N}, \quad j = 1, \dots, n, \quad (z_1, z_2, \dots) \in \Delta. \tag{2}$$

In this paper, $\mathsf{GEM}\,(\theta, \alpha)$ is characterized by its product moments, and sampling distributions of some statistics of a random sample $\boldsymbol{X}$ are obtained based on the characterization. The main statistics are the frequencies of categories in a sample in the order of appearance, the intervals between new categories and the number of distinct observations. Three theorems on the sampling distributions are stated in Section 2, and their proofs and related results are given in Sections 3–5.

For $\mathsf{GEM}\,(\theta, 0)$ the frequencies of categories in a sample from $\boldsymbol{Z}$ in the order of their sequence numbers were discussed by Donnelly and Joyce (1991), and Yamato and Nomachi (1997). The distribution of ordered frequencies of a sample in the order of appearance was discussed by Sibuya and Yamato (1995).

2 Main results on sampling distributions

For a random sample $\boldsymbol{X}= (X_1, \dots, X_n)$ from $\mathsf{GEM}\,(\theta, \alpha)$, the following statistics are defined. First, a sequence $(B_1, \dots, B_n)$ is defined by

$$B_1 := 1 \quad \text{and} \quad B_j := \prod_{i=1}^{j-1} \mathrm{I}[X_j \ne X_i], \quad j = 2, \dots, n, \tag{3}$$

where $\mathrm{I}[\,\cdot\,]$ is 1 if the bracketed event is true and 0 if false. That is, $B_j = 1$ if X_j is a new category, or a new number, and $= 0$ otherwise.

$$K_n := \sum_{j=1}^{n} B_j = 1 + \sum_{j=2}^{n} \mathrm{I}[X_j \neq X_k,\, k = 1, \ldots, j-1], \tag{4}$$

is the number of different categories of X. A realization $K_n = k$ is sometimes assumed without explicitly being mentioned.

$$\nu_1 := 1, \quad \text{and} \quad \nu_j := \min\left\{\ell\,;\, \sum_{i=1}^{\ell} B_i = j\right\}, \quad j = 2, \ldots, k,$$

is the 'time' when the j-th category appears first in $\boldsymbol{X}$, and the j-th category is denoted by

$$\widetilde{X_j} := X_{\nu_j}, \quad j = 1, \ldots, k.$$

Especially, $\widetilde{X_1} = X_1$. The number of components which are equal to $\widetilde{X_j}$ is denoted by

$$C_{nj} := \sum_{i=1}^{n} \mathrm{I}[X_i = \widetilde{X_j}], \quad j = 1, \ldots, k. \tag{5}$$

The number of components between new categories, or the 'waiting time' for a new category, is defined by

$$D_{nj} := \nu_{j+1} - \nu_j, \quad 1 \leq j < k, \quad \text{and} \quad D_{n,1} := n, \quad \text{if } k = 1. \tag{6}$$

That is, D_{n1} is the number of observations before $\widetilde{X_2}$, and D_{nj} is the number of observations between $\widetilde{X_j}$ and $\widetilde{X_{j+1}}$ including the former. If $k > 1$,

$$\text{`` } D_{nj} = d_j, \quad j = 1, \ldots, k-1\text{''} \quad \Leftrightarrow \quad \text{``}\widetilde{X_j} = X_{\nu_j} = X_{d_1+\cdots+d_{j-1}+1}, \quad j = 2, \ldots, k\text{''},$$

and $d_1 + \cdots + d_{k-1} < n$. The 'still waiting time' is similarly defined by

$$D_{nk} := n - \nu_k + 1, \quad \text{if} \quad k > 1, \quad \text{and} \quad D_{n1} := n,$$

as a convention.

The main results of this paper are as follows.

Theorem 1 *Let $\boldsymbol{Z}$ be a random vector of* GEM (θ, α), *and $\boldsymbol{X}$ be a sample of size n from $\boldsymbol{Z}$. The joint probability distribution of the number K_n of distinct observations $\widetilde{X_j}$ of $\boldsymbol{X}$, and the number C_{nj} (5) of observations which are the same as $\widetilde{X_j}$, $j = 1, \ldots, K_n$, in $\boldsymbol{X}$ is*

$$P(K_n = k,\, C_{n1} = c_1, \ldots, C_{nk} = c_k) = \frac{n!\theta^{[k:\alpha]}}{\theta^{[n]}} \prod_{i=1}^{k} \frac{(1-\alpha)^{[c_i-1]}}{(\sum_{j=i}^{k} c_j)(c_i - 1)!}, \tag{7}$$

where $1 \leq k \leq n$, $c_j > 0$, $j = 1, \ldots, k$; $c_1 + \cdots + c_k = n$; $x^{[j]} = x(x+1)\cdots(x+j-1)$ *and* $\theta^{[k:\alpha]} = \theta(\theta+\alpha)\cdots(\theta+(k-1)\alpha)$.

Conversely, let $\boldsymbol{Z}$ *be a random vector on* Δ, *and* $\boldsymbol{X}$ *be a random sample from* $\boldsymbol{Z}$. *If the components of* $\boldsymbol{Z}$ *do not tie a.s., and the joint distribution of* $(K_n, C_{n1}, \ldots, C_{nK_n})$ *has the distribution* (7), *then the size-biased permutation* $\boldsymbol{V}$ *of* $\boldsymbol{Z}$ *has* $\mathsf{GEM}\ (\theta, \alpha)$.

The distribution (7), in the case $\alpha = 0$, is known as DTG (Donnelly-Tavaré-Griffiths) sampling formula, and we call (7) as the *two-parameter DTG distribution*, and express it as $\mathsf{DTG}\ (\theta, \alpha)$.

In the following two theorems, statistics are assumed of a sample $\boldsymbol{X}$ of size n from a random $\mathsf{GEM}\ (\theta, \alpha)$ vector.

Theorem 2 *The joint distribution of* K_n *and* D_{nj}, $j = 1, \ldots, k-1$, (6), *is*

$$P(K_n = k,\ D_{n1} = d_1, \ldots, D_{n,k-1} = d_{k-1}) = \frac{\theta^{[k:\alpha]}}{\theta^{[n]}} \prod_{i=1}^{k} (1 - i\alpha + \sum_{j=1}^{i-1} d_j)^{[d_i - 1]}, \tag{8}$$

where $d_k = n - (d_1 + \cdots + d_{k-1}) > 0$.

For $\alpha = 0$, this distribution is called as DTG II formula (Yamato (1997)), and we term it the *two-parameter DTG II distribution*, and express it as $\mathsf{DTG\ II}\ (\theta, \alpha)$.

Theorem 3 *The number* K_n (4) *of different categories has the probability distribution,*

$$P(K_n = k) = \frac{\theta^{[k:\alpha]}}{\theta^{[n]}} \frac{c(n,k;\alpha)}{\alpha^k}, \quad k = 1, 2, \ldots, n, \tag{9}$$

where $c(n,k;\alpha) = (-1)^{n-k}\, C(n,k;\alpha)$, *and* $C(n,k;\alpha)$ *is the C-number of Charalambides and Singh* (1988).

This is a marginal distribution of (7) and (8). Pitman (1999) derived the distribution (9) of K_n by a different method.

The *C-number* is a type of generalized Stirling number defined by the polynomial identity

$$(st)^{(n)} = \sum_{k=1}^{n} C(n,k;s)\, t^{[k]}, \quad t^{(k)} = t\,(t-1)\ \cdots\ (t-k+1),$$

or equivalently $(st)^{[n]} = \sum_{k=1}^{n} c(n,k;s)\, t^{[k]}$. In terms of the unsigned Stirling number of the first kind $\left[{n \atop m}\right]$, and the Stirling number of the second kind $\left\{{n \atop m}\right\}$,

$$\frac{c(n,k;\alpha)}{\alpha^k} = \sum_{r=k}^{n} \begin{bmatrix} n \\ r \end{bmatrix} \begin{Bmatrix} r \\ k \end{Bmatrix} (-\alpha)^{r-k},$$

and $c(\cdot,\cdot;\cdot)$ satisfies the recurrence formula

$$c(n+1,k;s) = (n - k\,s)\, c(n,k;s) + s\, c(n,k-1;s), \tag{10}$$

with the initial conditions $c(0,0;s)=1$, $c(n,0;s)=0\,(n>0)$, and $c(0,k;s)=0\,(k>0)$, $\forall s$. See Pitman (1999) and Yamato and Sibuya (2000) for further discussions.

3 Two-parameter DTG distribution

The two-parameter GEM distribution is characterized as follows.

Proposition 1 *A random vector* $\boldsymbol{V}=(V_1,V_2,\dots)$ *on* Δ *has* $\mathsf{GEM}\ (\theta,\alpha)$ *if and only if*

$$E[V_1^{c_1-1}(1-V_1)V_2^{c_2-1}(1-V_1-V_2)V_3^{c_3-1}\cdots(1-V_1-\cdots-V_{k-1})V_k^{c_k-1}]$$

$$=\frac{\theta^{[k:\alpha]}}{\theta^{[c_1+\cdots+c_k]}}\prod_{i=1}^{k}(1-\alpha)^{[c_i-1]},\quad \forall k,c_1,\dots,c_k\in\mathcal{N}. \tag{11}$$

For $\mathsf{GEM}\ (\theta,0)$, this proposition was shown by Sibuya and Yamato (1995).

Proof. Recall that a beta distribution with any parameter is uniquely determined by its moments, and that for a random variable U of the beta distribution $\mathsf{Be}\ (a,b)$, $E[U^r(1-U)^s]=a^{[r]}b^{[s]}/(a+b)^{[r+s]}$. If $\boldsymbol{V}=(V_1,V_2,\dots)$ is a $\mathsf{GEM}\ (\theta,\alpha)$ random vector, then (11) is obtained directly from this moment and (1).

Conversely, assume (11), which is, with $k=1$, $E[V_1^{c_1-1}]=(1-\alpha)^{[c_1-1]}/(1+\theta)^{[c_1-1]}$ meaning that V_1 has $\mathsf{Be}\ (1-\alpha,\theta+\alpha)$. To advance the induction step from $k=j$ to $j+1$, assume that V_1,$V_2/(1-V_1)$, ... ,$V_j/(1-V_1-\cdots-V_{j-1})$ are independent and have $\mathsf{Be}\ (1-\alpha,\theta+\alpha)$, $\mathsf{Be}\ (1-\alpha,\theta+2\alpha)$, ... ,$\mathsf{Be}\ (1-\alpha,\theta+j\alpha)$, respectively. The expression (11) with $k=j+1$ can be written as

$$E[V_1^{c_1-1}(1-V_1)^{c_2+\cdots+c_{j+1}}(\frac{V_2}{1-V_1})^{c_2-1}(1-\frac{V_2}{1-V_1})^{c_3+\cdots+c_{j+1}}\cdots$$

$$(\frac{V_j}{1-V_1-\cdots-V_{j-1}})^{c_j-1}(1-\frac{V_j}{1-V_1-\cdots-V_{j-1}})^{c_{j+1}}(\frac{V_{j+1}}{1-V_1-\cdots-V_j})^{c_{j+1}-1}]$$

$$=E[V_1^{c_1-1}(1-V_1)^{c_2+\cdots+c_{j+1}}]E[(\frac{V_2}{1-V_1})^{c_2-1}(1-\frac{V_2}{1-V_1})^{c_3+\cdots+c_{j+1}}]\cdots$$

$$E[(\frac{V_j}{1-V_1-\cdots-V_{j-1}})^{c_j-1}(1-\frac{V_j}{1-V_1-\cdots-V_{j-1}})^{c_{j+1}}]E[B_{j+1}^{c_{j+1}-1}],$$

where $c_1,\dots,c_{j+1}$ are positive integers and B_{j+1} is a $\mathsf{Be}\ (1-\alpha,\theta+(j+1)\alpha)$ random variable. In the above relation, replace c_1 to c_1+l_1 and sum over $l_1=1,2,\dots$. Then after repeating these calculations $c_2+\cdots+c_{j+1}$ times, the term $(1-V_1)^{c_2+\cdots+c_{j+1}}$ disappears in both sides. Similarly for $i=2,\dots,j$, replace c_i by c_i+l_i and sum over $l_i=1,2,\dots$. Then after repeating these calculations $c_{i+1}+\cdots+c_{j+1}$ times, we obtain

$$E[V_1^{c_1-1}(\frac{V_2}{1-V_1})^{c_2-1}\cdots(\frac{V_j}{1-V_1-\cdots-V_{j-1}})^{c_j-1}(\frac{V_{j+1}}{1-V_1-\cdots-V_j})^{c_{j+1}-1}]$$

$$=E[V_1^{c_1-1}]E[(\frac{V_2}{1-V_1})^{c_2-1}]\cdots E[(\frac{V_j}{1-V_1-\cdots-V_{j-1}})^{c_j-1}]E[B_{j+1}^{c_{j+1}-1}],\quad \forall c_1,\dots,c_{j+1}\in\mathcal{N}.$$

Hence, V_1,$V_2/(1-V_1)$, $\dots$,$V_j/(1-V_1-\cdots-V_{j-1})$, $V_{j+1}/(1-V_1-\cdots-V_j)$ are independent and $V_{j+1}/(1-V_1-\cdots-V_j)$ has $\mathsf{Be}\,(1-\alpha,\theta+(j+1)\alpha)$, and the induction is complete. □

Let $\boldsymbol{X}$ be a sample of a random vector $\boldsymbol{Z}$ on Δ, and let $\boldsymbol{V}=(V_1,V_2,\dots)$ be the size-biased permutation of $\boldsymbol{Z}$. Because of (2) and (1), it is shown in Lemma 3 of Sibuya and Yamato (1995), that

$$P(K_n=k,\ X_1=\cdots=X_{c_1},X_{c_1+1}=\cdots=X_{c_1+c_2},\dots,X_{c_1+\cdots+c_{k-1}+1}=\cdots=X_{c_1+\cdots+c_k}) \tag{12}$$

$$=E[V_1^{c_1-1}(1-V_1)V_2^{c_2-1}(1-V_1-V_2)V_3^{c_3-1}\cdots(1-V_1-\cdots-V_{k-1})V_k^{c_k-1}],\quad \forall c_1,\dots,c_k\in\mathcal{N}.$$

Assume $\boldsymbol{V}$ has $\mathsf{GEM}\,(\theta,\alpha)$. Since $\mathsf{GEM}\,(\theta,\alpha)$ distribution is invariant under size-biased permutation (Pitman (1996a)), the assumption that $\boldsymbol{V}=(V_1,V_2,\dots)$ has $\mathsf{GEM}\,(\theta,\alpha)$ includes the case where $\boldsymbol{Z}$ itself is $\mathsf{GEM}\,(\theta,\alpha)$. Note that none of the components of $\boldsymbol{Z}$ tie almost surely if $W_1,W_2,\dots$ are independent and have continuous distributions (Gnedin(1998)).

Proposition 2 *For any integers k $(1\le k\le n)$ and $c_1,\dots,c_k\in\mathcal{N}$ satisfying $c_1+\cdots+c_k=n$,*

$$P(K_n=k,\ X_1=\cdots=X_{c_1},X_{c_1+1}=\cdots=X_{c_1+c_2},\dots,X_{c_1+\cdots+c_{k-1}+1}=\cdots=X_{c_1+\cdots+c_k})$$

$$=\frac{\theta^{[k:\alpha]}}{\theta^{[n]}}\prod_{i=1}^{k}(1-\alpha)^{[c_i-1]}. \tag{13}$$

Now we are ready to prove Theorem 1. The probability (13) depends only on k and $c_1,\dots,c_k$, and does not depend on $(\widetilde{X_1},\dots,\widetilde{X_n})$ nor arrangements of $(X_1,\dots,X_n)$. Hence, the number of arrangements such that $(C_{n1},\dots,C_{nk})=(c_1,\dots,c_k)$ is $\binom{n-1}{c_1-1}\binom{n-c_1-1}{c_2-1}\cdots\binom{n-c_1-\cdots-c_{k-1}-1}{c_k-1}$. Multiply (13) by the number of arrangements to get (7).

Conversely, suppose that a random vector $\boldsymbol{Z}$ on Δ has components which do not tie almost surely. Suppose that the statistics $(K_n,C_{n1},\dots,C_{nK_n})$ of a sample from $\boldsymbol{Z}$ has $\mathsf{DTG}\,(\theta,\alpha)$. It means that (13) is valid, hence (12) and (11) are also valid. Hence, by Proposition 1, the size-biased permutation has $\mathsf{GEM}\,(\theta,\alpha)$, and Theorem 1 is proved.

The results stated in Proposition 1, Proposition 2 and Theorem 1 are stated in formulae (7) and (8) of Pitman (1995) in a different context.

4 Intervals between new observations

Because of the independence of the sample $\boldsymbol{X}$ given $\boldsymbol{Z}$, from (13),

$$P(K_{n+1}=k,\ X_1=\cdots=X_{c_1},\dots,X_{n-c_k+1}=\cdots=X_n,X_{n+1}=X_{c_1+\cdots+c_j})$$

$$=\frac{c_j-\alpha}{\theta+n}P(K_n=k,\ X_1=\cdots=X_{c_1},X_{c_1+1}=\cdots=X_{c_1+c_2},\dots,X_{n-c_k+1}=\cdots=X_n),\quad 1\le j\le k\le n.$$

Hence,

$$P(X_{n+1} = \widetilde{X_j} \mid K_n = k,\ X_1 = \cdots = X_{c_1}, X_{c_1+1} = \cdots = X_{c_1+c_2}, \ldots, X_{n-c_k+1} = \cdots = X_n)$$

$$= (c_j - \alpha)/(\theta + n), \quad 1 \leq j \leq k \leq n, \tag{14}$$

$$P(X_{n+1} \neq X_1, \ldots, X_n \mid K_n = k,\ X_1 = \cdots = X_{c_1}, X_{c_1+1} = \cdots = X_{c_1+c_2}, \ldots,$$

$$X_{n-c_k+1} = \cdots = X_n) = (\theta + k\alpha)/(\theta + n), \quad 1 \leq k \leq n. \tag{15}$$

The sequence $X_1, X_2, \ldots$ as equivalent class of $\widetilde{X_j},\ j = 1, 2, \ldots,$ is generated by an urn model, Pitman (1995, 1996b).

Consider the random binary sequence $B_1, B_2, \ldots,$ (3). Since the conditional probabilities (14) and (15) depend only on the number $K_n = k$ of distinct observations,

$$P(B_{j+1} = 1 \mid B_1 = b_1, \ldots, B_j = b_j) = \frac{\theta + (b_1 + \cdots + b_j)\alpha}{j + \theta}, \tag{16}$$

and

$$P(B_{j+1} = 0 \mid B_1 = b_1, \ldots, B_j = b_j) = \frac{j - (b_1 + \cdots + b_j)\alpha}{j + \theta}, \quad j = 1, 2, \ldots \tag{17}$$

For $1 \leq k \leq n$ and $d_1, \ldots, d_{k-1} \in \mathcal{N}$,

$$P(K_n = k, D_{n1} = d_1, \ldots, D_{nk} = d_k) = P(B_1 = 1, B_2 = \cdots = B_{d_1} = 0, B_{d_1+1} = 1, B_{d_1+2} = \cdots$$

$$= B_{d_1+d_2} = 0, B_{d_1+d_2+1} = 1, \ldots, B_{d_1+\cdots+d_{k-1}+1} = 1, B_{d_1+\cdots+d_{k-1}+2} = \cdots = B_{d_1+\cdots+d_k} = 0).$$

Hence, (8), or Theorem 2, is proved.

The following propositions deal with the marginal distributions of DTG (θ, α).

Proposition 3 *For r such that $1 \leq r < n$ and for $d_1, \ldots, d_r \in \mathcal{N}$,*

$$P(D_{n1} = d_1, \ldots, D_{nr} = d_r) = \frac{\theta^{[r+1:\alpha]}}{\theta^{[d(r)+1]}} \prod_{i=1}^{r} (1 - i\alpha + \sum_{j=1}^{i-1} d_j)^{[d_i - 1]}, \tag{18}$$

provided that $d(r) := d_1 + \cdots + d_r < n$.

Proof: For $d_1 + \cdots + d_r < n$, we have

$$P(D_{n1} = d_1, \ldots, D_{nr} = d_r) = P(B_1 = 1, B_2 = \cdots = B_{d_1} = 0, B_{d_1+1} = 1, \ldots,$$

$$B_{d(r-1)+1} = 1, B_{d(r-1)+2} = \cdots = B_{d(r)} = 0, B_{d(r)+1} = 1).$$

Hence (18) is obtained using (16) and (17). □

If $k = 1$, $D_{n1} = n$ by convention of (8), and

$$P\{D_{n1} = n\} = (1-\alpha)^{[n-1]} \Big/ (1+\theta)^{[n-1]},$$

from (16) and (17). Otherwise, $k > 1$ or $D_{n1} < n$, as a special case of (18),

$$P\{D_{n1} = d_1\} = (\theta+\alpha)(1-\alpha)^{[d_1-1]} \Big/ (1+\theta)^{[d_1-1]}.$$

Hence, $D_{n1} - 1$ has the bounded Waring distribution BWa $(n-1, 1+\theta, 1-\alpha)$ (see Yamato (1997)). The following proposition extends this result.

Proposition 4 *Let $r, d_1, \ldots, d_{r-1} \in \mathcal{N}$ be such that $1 < r \le n$ and $d(r-1) = d_1 + \cdots + d_{r-1} < n$. If $r < k$,*

$$P(D_{nr} = d_r \mid D_{n1} = d_1, \ldots, D_{n,r-1} = d_{r-1}) = \frac{(\theta + r\alpha)(d(r-1)+1-r\alpha)^{[d_r-1]}}{(d(r-1)+1+\theta)^{[d_r]}}, \quad 1 \le d_r < d(r-1), \tag{19}$$

and if $r = k$ or $d_r = n - d(r-1)$,

$$P(D_{nr} = d_r \mid D_{n1} = d_1, \ldots, D_{n,r-1} = d_{r-1}) = \frac{(d(r-1)+1-r\alpha)^{[d_r-1]}}{(d(r-1)+1+\theta)^{[d_r-1]}}. \tag{20}$$

Proof: For $d_r = 1, 2, \ldots, n - d(r-1) - 1$, dividing (18) by the equation with $r-1$ instead of r we have (19). For $d_r = n - d(r-1)$, dividing (8) with $k = r$ by (18) with $r-1$ instead of r we have (20). □
The probabilities (19) and (20) show that the conditional distribution of $D_{nr} - 1$ given $d(r-1)$ is the bounded Waring distribution BWa $(n - d(r-1) - 1, d(r-1) + 1 + \theta, d(r-1) + 1 - r\alpha)$.

5 Number of distinct observations

The distribution of K_n (4) for $\alpha = 0$ is $\mathsf{STR1F}$ (n, θ), a subfamily of the Stirling family of probability distributions, defined by $P(K_n = k) = \left[{n \atop k}\right]\theta^k/\theta^{[n]}$ $k = 1, \ldots, n$, Sibuya (1988). The generalized distribution (9) of K_n in Theorem 3 is proved as follows.

First, special cases of (9) are shown: By definition $B_1 = K_1 = 1$. From the conditional probabilities (16) and (17), $P(K_2 = 1) = P(B_1 = 1, B_2 = 0) = (1-\alpha)/(\theta+1)$ and $P(K_2 = 2) = P(B_1 = B_2 = 1) = (\theta+\alpha)/(\theta+1)$. Similarly, $P(K_n = 1) = P(B_1 = 1, B_2 = \cdots = B_n = 0) = \theta(1-\alpha)^{[n-1]}/\theta^{[n]}$ and $P(K_n = n) = P(B_1 = \cdots = B_n = 1) = \theta^{[n:\alpha]}/\theta^{[n]}$.

Assume (9) for a positive integer n, and from (16) and (17),

$$P(K_{n+1} = k+1 \mid K_n = k) = \frac{\theta + k\alpha}{\theta + n}, \quad P(K_{n+1} = k \mid K_n = k) = \frac{n - k\alpha}{\theta + n}. \tag{21}$$

This relation and

$$P(K_{n+1}=k)=P(K_{n+1}=k \mid K_n=k)P(K_n=k)+P(K_{n+1}=k \mid K_n=k-1)P(K_n=k-1)$$

show that $\alpha^k\theta^{[n]}P(K_n=k)/\theta^{[k:n]}$ satisfies the recurrence formula (10). Hence, (9) and Theorem 3 are proved. □

On the right-hand side of (21), we have $n-k\alpha>0$ for $k=1,\dots,n$ because of $0\le\alpha<1$. Since $c(n,n;\alpha)=\alpha^n$ and $c(n,0;\alpha)=0\,(n>0)$, using (21) it is seen by induction that $c(n,k;\alpha)>0$ for $k=1,\dots,n$. Similarly, using (21) it is derived by induction that $c(n,k;1/2)=2^{k-2n}(2n-k-1)!/[(n-k)!(k-1)!]$, $k=1,\dots,n$.

Thus for $\theta=0$ and $\alpha=1/2$, we have

$$P(K_n=k)=\left(\begin{array}{c} 2n-k-1 \\ n-1 \end{array}\right)2^{k+1-2n},$$

and for $\theta=\alpha=1/2$,

$$P(K_n=k)=\frac{k(n-1)!}{(3/2)^{[n-1]}}\left(\begin{array}{c} 2n-k-1 \\ n-1 \end{array}\right)2^{k+1-2n},$$

(Pitman(1997), p.81). Using the descending factorial, the probability (9) can be written as

$$P(K_n=k)=\frac{(-\theta/\alpha)^{(k)}}{(-\theta)^{(n)}}C(n,k;\alpha),\quad k=1,\dots,n.$$

This is the same form as the probability (6.6) of Charalambides and Singh (1988), p.2564, which was derived in a different context.

For $(C_{n1},\dots,C_{nk})$ with $K_n=k$, define

$$M_j:=\sum_{i=1}^{n}\mathrm{I}[C_{ni}=j],\quad j=1,\dots,n,$$

which is the number of categories with frequency j, sometimes called 'frequency of frequencies'. Note that $\sum_{j=1}^n M_j=k$ and $\sum_{j=1}^n j\,M_j=n$. For a fixed $(m_1,\dots,m_n)$ satisfying $\sum_{j=1}^n m_j=k$ and $\sum_{j=1}^n j\,m_j=n$, multiply (13) by the number of ways which gives $(m_1,\dots,m_n)$ to get the Pitman sampling formula, Pitman (1995),

$$P(K_n=k,(M_1,\dots,M_n)=(m_1,\dots,m_n))=n!(\theta^{[k:\alpha]}/\theta^{[n]})\prod_{j=1}^{n}\{((1-\alpha)^{[j-1]}/j!)^{m_j}/m_j!\}$$

$$=\frac{(-1)^{n-k}\theta^{[k:\alpha]}n!}{\alpha^k\theta^{[n]}}\prod_{j=1}^{n}\frac{1}{m_j!}\binom{\alpha}{j}^{m_j}. \tag{22}$$

By the summation $\sum_1$ over $m_1, \ldots, m_n \in \mathcal{N}$ satisfying $\sum_{j=1}^n m_j = k$ and $\sum_{j=1}^n j\, m_j = n$ with fixed k and n, $P(K_n = k) = \sum_1 P(K_n = k, (M_1, \ldots, M_n) = (m_1, \ldots, m_n))$. Applying to this summation an expression of C-number given by Charalambides and Singh (1988), p.2553,

$$C(n, k; \alpha) = \sum\nolimits_1 \frac{n!}{m_1! \cdots m_n!} \binom{\alpha}{1}^{m_1} \cdots \binom{\alpha}{n}^{m_n},$$

we can also get the distribution (9) of K_n.

Acknowledgment

The authors would like to thank Professor Jim Pitman for his comments.

References

[1] Charalambides, Ch.A. and Singh, J. (1988), A review of the Stirling numbers, their generalizations and statistical applications, *Commun. Statist.-Theory Meth.* **17**, 2533–2595.

[2] Donnelly, P. and Joyce, P. (1991), Consistent ordered sampling distribution: Characterization and convergence, *Adv. Appl. Probab.* **23**, 229–258

[3] Gnedin, A.V. (1998), On convergence and extensions of size-biased permutation, *J. Appl. Prob.* **35**, 642–650.

[4] Johnson, N.L., Kotz, S. and Balakrishnan, N.B. (1997), *Discrete Multivariate Distributions*, Wiley, New York.

[5] Patil, G.P. and Taillie, C. (1977), Diversity as a concept and its implications for random communities, *Bull. Internat. Statist. Inst.* **47**, 497–515.

[6] Pitman, J. (1995), Exchangeable and partially exchangeable random partitions, *Probab. Theory Relat. Fields* **102**, 145–158.

[7] Pitman, J. (1996a), Random discrete distributions invariant under size-biased permutation, *Adv. Appl. Prob.* **28**, 525–539.

[8] Pitman, J. (1996b), Some developments of the Blackwell-MacQueen urn scheme, in: T.S. Ferguson, L.S. Shapley and J.B. MacQueen eds., *Statistics, Probability and Game Theory. Papers in Honor of David Blackwell*, IMS, Hayward, CA, 245–267.

[9] Pitman, J. (1997), Partition structures derived from Brownian motion and stable subordinaters, *Bernoulli.* **3**, 79–96.

[10] Pitman, J. (1999), Brownian motion, bridge, excursion, and meander characterized by sampling at independent uniform times, *Electronic J. Probability*, **4**, Paper no. 11, 1–33.

[11] Pitman, J. and M. Yor (1997), The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator, *Ann. Probab.* **25**, 855–900.

[12] Sibuya, M. (1988), Log-concavity of Stirling numbers and unimodality of Stirling distributions, *Ann. Inst. Statist. Math.*. **40**, 693–714.

[13] Sibuya, M. and Yamato, H. (1995), Ordered and unordered random partitions of an integer and the GEM distribution, *Statist. Prob. Letters.* **25**, 177–183.

[14] Yamato, H. (1997), On the Donnelly-Tavaré-Griffiths formula associated with coalescent, *Commun. Statist.-Theory Meth.* **26**, 589–599.

[15] Yamato, H. and Nomachi, T. (1997), The distribution of frequencies of discrete order statistics and the Donnelley-Tavaré-Griffiths formula, *J. Nonparametric Statist.* **8**, 355–363.

[16] Yamato, H. and Sibuya, M. (2000), Moments of some statistics of Pitman sampling formula, *Bulletin of Informatics and Cybernetics* (Fukuoka) **32**, 1–10.