ON NUMBER OF UNOBSERVED POSITIVE INTEGERS LESS THAN THE MAXIMUM SAMPLE FROM GEM DISTRIBUTION

Yamato, Hajime
Emeritus of Kagoshima University

https://doi.org/10.5109/4844361

出版情報：Bulletin of informatics and cybernetics. 54 (5), pp.1-13, 2022. 統計科学研究会
バージョン：
権利関係：

ON NUMBER OF UNOBSERVED POSITIVE INTEGERS LESS THAN THE MAXIMUM SAMPLE FROM GEM DISTRIBUTION

by

Hajime Yamato

Reprinted from the Bulletin of Informatics and Cybernetics
Research Association of Statistical Sciences, Vol.54, No. 5

FUKUOKA, JAPAN
2022

Bulletin of Informatics and Cybernetics, Vol. 54, No. 5, 2022

ON NUMBER OF UNOBSERVED POSITIVE INTEGERS LESS THAN THE MAXIMUM SAMPLE FROM GEM DISTRIBUTION

By

Hajime Yamato*

Abstract

For a sample from GEM distribution considered as a random discrete distribution of positive integers, there are probably unobserved less positive integers than the maximum sample. The number of these integers follows a mixed Poisson distribution. Here, we provide a simple proof of the convergence in distribution of this number. Similarly, its asymptotic distribution is a mixed Poisson. We derived the upper bounds for the total variation between the finite and asymptotic distributions. These distributions are illustrated by examples; the related total values are provided in table 1.

Key Words and Phrases: convergence in distribution, GEM distribution, mixed Poisson distribution, mixing distribution, number of unobserved positive integers, order statistics, total variation distance.

1. Introduction

Let $\xi_1, \xi_2, \dots$ be independent and identically distributed random variables having the beta distribution $Beta(1, \theta)$ with density $\theta(1-x)^{\theta-1}$ $(0 < x < 1,\ \theta > 0)$. For the following random frequencies $(P_1, P_2, \dots)$:

$$P_1 = \xi_1, \quad P_j = (1-\xi_1)\cdots(1-\xi_{j-1})\xi_j \ \ (j = 2, 3, \dots). \tag{1}$$

The distribution of $\mathbf{P} = (P_1, P_2, \dots)$ is well-known as a GEM distribution with parameter θ. The term GEM was named after Griffith (unpublished notes), Engen (1975) and McCloskey (1965) by Ewens (1990, p.217). The GEM distribution is used in many fields such as Bayesian statistics, genetics, ecology, etc. (Johnson et al.,1997)). For a sample of size n from $\mathbf{P}$, if C_l is the number of values exactly observed l times for $l = 1, 2, \dots, n$. The distribution of $(C_1, C_2, \dots, C_n)$ is known as Ewens' sampling formula: $P\big((C_1, C_2, \dots, C_n) = (c_1, c_2, \dots, c_n)\big) = \theta^k n!/[(\theta)_n \prod_{j=1}^{k} j^{c_j} c_j!]$, where $k\ (= c_1 + \cdots + c_n)$ is the number of distinct values in the sample and $(\theta)_n = \theta(\theta+1)\dots(\theta+n-1)$ (Ewens,1972, p.88; Antoniak,1974, p.1162).

In this paper, $\mathbf{P} = (P_1, P_2, \dots)$ is considered as a random discrete distribution of a set $\mathbf{N}$ of positive integers. For the $\mathbf{P}$ sample, there are probably less unobserved positive integers than the maximum in the observed sample. We aimed to determine the number $K_{0:n}$ of unobserved positive integers lower than the maximum. This is equivalent to the

* Emeritus of Kagoshima University, Take 3-32-1-708, Kagoshima 890–0045, Japan

number of empty boxes in Gnedin, Iksanov, Negadajlov and Rösler (2009) and Gnedin, Iksanov and Marynych (2010). We consider $K_{0:n}$ based on the order statistics for a sample of size n from $\mathbf{P}$ on $\mathbf{N}$ having a GEM distribution with parameter $\theta(>0)$. We express the order statistics as $X_{1:n} \leq \cdots \leq X_{n:n}$ and put their differences as follows:

$$\widehat{G}_{n:n} = X_{1:n} - 1, \quad \widehat{G}_{i:n} = X_{n-i+1:n} - X_{n-i:n} \ \ (i = 1, \ldots, n-1).$$

LEMMA 1.1. (Pitman and Yakubovich (2017, Theorem 1.1.)) *$\widehat{G}_{1:n}, \widehat{G}_{2:n}, \ldots, \widehat{G}_{n:n}$ are independent. For $i = 1, 2, \ldots, n$, $\widehat{G}_{i:n}$ has the geometric distribution $Ge(p_i)$ ($p_i = i/(i+\theta)$) whose probability function is $P(\widehat{G}_{i:n} = k) = p_i(1-p_i)^k$ ($k = 0, 1, 2, \ldots$).*

This lemma is explained in Appendix 3.1. $K_{0:n}$ is the number of unobserved positive integers less than $X_{n:n}$. We put $u_+ = u$ ($u \geq 0$), 0 ($u < 0$). Then, $K_{0:n}$ is calculated by the sum of $\widehat{G}_{n:n}$ and $(\widehat{G}_{i:n} - 1)_+$ ($i = 1, \ldots, n-1$) as follows:

$$K_{0:n} = \widetilde{G}_{1:n-1} + \widehat{G}_{n:n}, \ \ \widetilde{G}_{1:n-1} = (\widehat{G}_{1:n} - 1)_+ + (\widehat{G}_{2:n} - 1)_+ + \cdots + (\widehat{G}_{n-1:n} - 1)_+, \quad (2)$$

(see, Pitman and Yakubovich (2017, 1.14) and Appendix 3.2).

Let ξ be a random variable having the beta distribution $Beta(1, \theta)$. We consider $K_{0:\infty}$ as a random variable having a mixed Poisson distribution $Po(-\theta \log \xi)$. That is, $K_{0:\infty}$ has the Poisson distribution with the parameter $-\theta \log \xi$, given ξ. The random parameter of the mixed Poisson distribution is also called mixing parameter. For mixed Poisson distributions, see Johnson et al. (2005) and Grandell (1997).

Gnedin et al. (2009, p.1645) provided the relation equivalent to (2), showing the following Lemma 1.2 using a generating function, based on Bernoulli sieve different from our model.

LEMMA 1.2. (Gnedin, Iksanov, Negadajlov and Rösler (2009, Proposition 5.1) and Gnedin, Iksanov and Marynych (2010, Theorem 1.1.))

$K_{0:n} \xrightarrow{d} K_{0:\infty}$, that is, $K_{0:n}$ converges in distribution to $K_{0:\infty}$ as $n \to \infty$.

Instead of a generating function, we can show Lemma 1.2 using the expression $K_{0:n}$ and $K_{0:\infty}$ distributions as mixed Poisson distributions, as shown in section 2. In addition, we considered the upper bound of the total variation distance (TVD) $d_{TV}(K_{0:n}, K_{0:\infty})$ between $K_{0:n}$ and $K_{0:\infty}$, which is defined by

$$d_{TV}(K_{0:n}, K_{0:\infty}) = \frac{1}{2} \sum_{k=0}^{\infty} \big| P(K_{0:n} = k) - P(K_{0:\infty} = k) \big|.$$

This bound provides the convergence rate $K_{0:n}$ to $K_{0:\infty}$ as $n \to \infty$. Examples of the probability functions of $K_{0:n}$ and $K_{0:\infty}$ are illustrated for several θ's using the programming language R. The approximate values of $d_{TV}(K_{0:n}, K_{0:\infty})$ corresponding to these examples are provided together with the related total variation distance, in table 1. In section 3, the Appendix, we explain Lemma 1.1, the relation(2) and that the distribution of $(\widehat{G}_{j:n} - 1)_+$ is described by the mixed Poisson distribution. In addition, we provide an elementary proof of the relationship including Pitman and Yakubovich (2017; Proposition 7.1, (7.4)).

REMARK. Let $Q_0 = 1$ and $Q_j = \prod_{i=1}^{j}(1-\xi_i)$ $(j = 1, 2, \dots)$. Consider the GEM distribution expressed as $(0, \dots, Q_2, Q_1, Q_0)$ and set $S_0 = 0$ and $S_j = -\log Q_j$. Then, $0 < S_1 < S_2 < \cdots \uparrow \infty$ almost surely (a.s.). The point process based on $S_1, S_2, \dots$ is the Poisson process of intensity θ (see, Appendix 3.1). Gnedin, Iksanov, Negadajlov and Rösler (2009) and Gnedin, Iksanov and Marynych (2010) consider random exponential variables as the sample and the numbers of observations dropped into the intervals of the Poisson process.

2. Distribution and total variation distance

2.1. Distribution and convergence of $K_{0:n}$

Let $B_{q_j}(j = 1, 2, \dots)$ be independent random variables having the Bernoulli distribution $Ber(q_j)$ $(q_j = \theta/(j+\theta))$ and ε_j $(j = 1, 2, \dots)$ independent random variables having the exponential distribution $Exp(1)$. Let $\{B_{q_j}\}$ and $\{\varepsilon_j\}$ be mutually independent and

$$\Lambda_{n:\infty} = \sum_{j=n}^{\infty} B_{q_j} \frac{\varepsilon_j}{j} \quad (n = 1, 2, \dots).$$

Pitman and Yakubovich (2017, p.20) showed $(\widehat{G}_{j:n} - 1)_+$ has the mixed Poisson distribution $Po\big(\theta B_{q_j}\varepsilon_j/j\big)$ $(j = 1, \dots, n-1)$ and provided the following relation in the distribution (see Appendix A.9):

$$\Lambda_{1:\infty}\Big(= \sum_{j=1}^{\infty} B_{q_j} \frac{\varepsilon_j}{j}\Big) \stackrel{d}{=} -\log \xi, \tag{3}$$

where $\stackrel{d}{=}$ represents equivalent distribution. For the distribution of $(\widehat{G}_{j:n} - 1)_+$ and relation (3), see Appendix 3.3 and 3.5, respectively.

From Lemma 1.1, $\widehat{G}_{1:n}, \widehat{G}_{2:n}, \dots, \widehat{G}_{n-1:n}$ are independent. According to the above, $(\widehat{G}_{j:n} - 1)_+$ has a mixed Poisson distribution $Po\big(\theta B_{q_j}\varepsilon_j/j\big)$ for $j = 1, 2, \dots, n-1$. Therefore, as indicated in the Corollary A.2 of Appendix, $\widetilde{G}_{1:n-1} = (\widehat{G}_{1:n} - 1)_+ + \cdots + (\widehat{G}_{n-1:n} - 1)_+$ has the mixed Poisson distribution $Po(\theta\Lambda_{1:n-1})$, where

$$\Lambda_{1:n-1} = \sum_{j=1}^{n-1} B_{q_j} \frac{\varepsilon_j}{j} \ (n = 2, 3, \dots). \tag{4}$$

From Lemma 1.1, $\widehat{G}_{n:n}$ is independent of $\widetilde{G}_{1:n-1}(\sim Po(\theta\Lambda_{1:n-1}))$ and has the geometric distribution $Ge(n/(n+\theta))$. This distribution is equivalent to the mixed Poisson distribution $Po(\theta\varepsilon_n/n)$. Since ε_n and $\Lambda_{1:n-1}$ are independent, according to Proposition A.1 $K_{0:n} = \widetilde{G}_{1:n-1} + \widehat{G}_{n:n}$ has the following mixed Poisson distribution:

PROPOSITION 2.1.

$$K_{0:n} \ \sim \ Po\bigg(\theta\Lambda_{1:n-1} + \theta\frac{\varepsilon_n}{n}\bigg). \tag{5}$$

We prove Lemma 1.2 using the right-hand side expression. That is, we show the convergence of mixed variable $K_{0:n}$ through the convergence of the mixing variable

$\theta\Lambda_{n-1} + \theta\varepsilon_n/n$. Although this proof differs slightly from Yamato (2020, p.62), the authors consider it highly relevant.

Proof of Lemma 1.2. We consider the expectation of the absolute difference between the mixing parameters of $K_{0:\infty}$ and $K_{0:n}$. For $n > 1$, as $\Lambda_{n:\infty} = \Lambda_{1:\infty} - \Lambda_{1:n-1}$ we have the following:

$$\begin{aligned} E\left|\theta\Lambda_{1:\infty} - \theta\Big(\Lambda_{1:n-1} + \frac{\varepsilon_n}{n}\Big)\right| &= E\left|\theta\Lambda_{n:\infty} - \theta\frac{\varepsilon_n}{n}\right| \quad (6)\\ &\le E\Big[\theta\Big(\sum_{j=n+1}^{\infty} B_{q_j}\frac{\varepsilon_j}{j} + \Big(1 - B_{q_n}\Big)\frac{\varepsilon_n}{n}\Big)\Big]\\ &= \theta\Big(\sum_{j=n+1}^{\infty}\frac{\theta}{(j+\theta)j} + \frac{1}{n+\theta}\Big) \quad (7)\\ &\le \theta^2\sum_{j=n+1}^{\infty}\frac{1}{j^2} + \frac{\theta}{n+\theta} \le \frac{\theta^2}{n} + \frac{\theta}{n+\theta}, \quad (8)\end{aligned}$$

which converges to zero as $n \to \infty$. Thus, $\theta(\Lambda_{1:n-1} + \varepsilon_n/n)$'s probability converges to $\theta\Lambda_{1:\infty}$ $(n \to \infty)$. Therefore, $\theta\Lambda_{1:n-1} + \theta\varepsilon_n/n \xrightarrow{d} \theta\Lambda_{1:\infty}$. The convergence in distribution of the mixing parameter is equivalent to that of the mixed Poisson variable (see Grandell,1997; p.16). Thus, based on (5), we have $K_{0:n} \xrightarrow{d} Po(\theta\Lambda_{1:\infty})$. According to (3), this is equivalent to $K_{0:n} \xrightarrow{d} Po(-\theta\log\xi)$, which is equivalent to $K_{0:n} \xrightarrow{d} K_{0:\infty}$. □

The probability function (p.f.) of $K_{0:\infty}$ is shown in Appendix A.13, for $\theta > 0$. If θ is a positive integer, then the p.f. of $K_{0:\infty}$ is indicted in by Appendix A.14. For example,

$$P(K_{0:\infty} = k) = \frac{1}{2^{k+1}}\ (\theta = 1),\ P(K_{0:\infty} = k) = 3^{k+1}\Big(\frac{1}{4^{k+1}} - \frac{2}{5^{k+1}} + \frac{1}{6^{k+1}}\Big)\ (\theta = 3).$$

2.2. Total variation distance between $K_{0:n}$ and $K_{0:\infty}$

Let $\widetilde{G}_{n:\infty}$ be the random variable having the mixed Poisson distribution $Po(\theta\Lambda_{n:\infty})$ and independent of $\widetilde{G}_{1:n-1}$ $(\sim Po(\theta\Lambda_{1:n-1}))$. As $\Lambda_{1:n-1}$ and $\Lambda_{n:\infty}$ are independent, based on Corollary A.2, $\widetilde{G}_{1:n-1}+\widetilde{G}_{n:\infty}$ has the mixed Poisson distribution $Po(\theta(\Lambda_{1:n-1}+\Lambda_{n:\infty}))$. On the other hand, $K_{0:\infty} \sim Po(\theta\Lambda_{1:\infty})$, where $\Lambda_{1:\infty} = \Lambda_{1:n-1} + \Lambda_{n:\infty}$. Thus, we have

$$K_{0:\infty} \overset{d}{=} \widetilde{G}_{1:n-1} + \widetilde{G}_{n:\infty}.$$

The TVD between $K_{0:n}$ and $K_{0:\infty}$ is less than or equal to TVD between $\widehat{G}_{n:n}$ and $\widetilde{G}_{n:\infty}$ as shown below:

LEMMA 2.2.

$$d_{TV}(K_{0:n}, K_{0:\infty}) \le d_{TV}(\widehat{G}_{n:n}, \widetilde{G}_{n:\infty}). \quad (9)$$

Proof. Because $K_{0:n} = \widetilde{G}_{1:n-1} + \widehat{G}_{n:n}$ and $K_{0:\infty} \stackrel{d}{=} \widetilde{G}_{1:n-1} + \widetilde{G}_{n:\infty}$, we have

$$\begin{aligned} d_{TV}(K_{0:n}, K_{0:\infty}) &= \frac{1}{2}\sum_{k=0}^{\infty} \big|P(K_{0:n}=k) - P(K_{0:\infty}=k)\big| \\ &= \frac{1}{2}\sum_{k=0}^{\infty} \Big|\sum_{j=0}^{k} P(\widetilde{G}_{1:n-1}=k-j)\big[P(\widehat{G}_{n:n}=j) - P(\widetilde{G}_{n:\infty}=j)\big]\Big| \\ &\le \frac{1}{2}\sum_{j=0}^{\infty}\sum_{k=j}^{\infty} P(\widetilde{G}_{1:n-1}=k-j)\big|P(\widehat{G}_{n:n}=j) - P(\widetilde{G}_{n:\infty}=j)\big| \\ &= \frac{1}{2}\sum_{j=0}^{\infty} \big|P(\widehat{G}_{n:n}=j) - P(\widetilde{G}_{n:\infty}=j)\big|. \end{aligned}$$

Thus we get (9). □

LEMMA 2.3.

$$d_{TV}(\widehat{G}_{n:n}, \widetilde{G}_{n:\infty}) \le \min\left\{1, \frac{\theta^2}{n} + \frac{\theta}{n+\theta}\right\}. \tag{10}$$

Proof. To calculate the TVD among $\widehat{G}_{n:n}$ and $\widetilde{G}_{n:\infty}$, we consider their conditional distributions as follows:

$$\begin{aligned} d_{TV}(\widehat{G}_{n:n}, \widetilde{G}_{n:\infty}) &= \frac{1}{2}\sum_{k=0}^{\infty} \big|P(\widehat{G}_{n:n})=k) - P(\widetilde{G}_{n:\infty}=k)\big| \\ &= \frac{1}{2}\sum_{k=0}^{\infty} \big|E\big[P(\widehat{G}_{n:n})=k \mid \varepsilon_n) - P(\widetilde{G}_{n:\infty}=k \mid \Lambda_{n:\infty})\big]\big| \\ &\le E\Big[\frac{1}{2}\sum_{k=0}^{\infty} \big|P(\widehat{G}_{n:n})=k \mid \varepsilon_n) - P(\widetilde{G}_{n:\infty}=k \mid \Lambda_{n:\infty})\big|\Big]. \end{aligned} \tag{11}$$

The TVD between two Poisson distributions is less than or equal to the absolute difference of their parameters (Freedman (1974, p.260)). Given ε_n and $\Lambda_{n:\infty}$, $\widehat{G}_{n:n}$ and $\widetilde{G}_{n:\infty}$ have the Poisson distributions $Po(\theta\varepsilon_n/n)$ and $Po(\theta\Lambda_{n:\infty})$, respectively. Hence, based on (11) we have

$$d_{TV}(\widehat{G}_{n:n}, \widetilde{G}_{n:\infty}) \le E\left|\,\theta\frac{\varepsilon_n}{n} - \theta\Lambda_{n:\infty}\,\right|. \tag{12}$$

Using equations (6) and (8) on the right-hand side of equation (12), $d_{TV}(\widehat{G}_{n:n}, \widetilde{G}_{n:\infty})$ is less than or equal to the last term of equation (8). As the TVD is less than or equal to 1, we get equation (10). □

Based on Lemmas 2.2 and 2.3, we have the following:

THEOREM 2.4.

$$d_{TV}(K_{0:n}, K_{0:\infty}) \le \min\left\{1, \frac{\theta^2}{n} + \frac{\theta}{n+\theta}\right\}. \tag{13}$$

Here, we assume $\theta \geq 1$ and let $l := [\theta]$, where $[\ \ ]$ is the Gauss symbol. Using the relations $j + l \leq j + \theta$ and $\sum_{j=n+1}^{\infty} 1/[(j+l)j] = \{1/(n+1) + \cdots + 1/(n+l)\}/l$ to equations (6), (7), and (12), we have

$$d_{TV}(\widehat{G}_{n:n}, \widetilde{G}_{n:\infty}) \leq \min\left\{1, \frac{\theta^2}{l}\Big(\frac{1}{n+1} + \cdots + \frac{1}{n+l}\Big) + \frac{\theta}{n+\theta}\right\} \quad (l = [\theta],\, \theta \geq 1). \tag{14}$$

For $\theta \geq 1$, $\sum_{j=n+1}^{\infty} 1/[(j+l)j] < \sum_{j=n+1}^{\infty} 1/j^2$ since $l \geq 1$. Therefore, the right-hand side of equation (14) is less than the right-hand side of equation (10) for $\theta \geq 1$. For example, for $\theta = 1$ and $\theta = 2$, (14) gives

$$d_{TV}(\widehat{G}_{n:n}, \widetilde{G}_{n:\infty}) \leq \frac{2}{n+1}\ (\theta = 1, n > 1),$$

$$d_{TV}(\widehat{G}_{n:n}, \widetilde{G}_{n:\infty}) \leq \frac{2}{n+1} + \frac{4}{n+2}\ (\theta = 2, n > 4). \tag{15}$$

The right-hand sides of these inequalities are less than the right-hand sides of equation (10), which are $1/n + 1/(n+1)$ $(\theta = 1)$ and $4/n + 2/(n+2)$ $(\theta = 2)$, respectively. Thus, with equations (9) and (14), we obtain a better evaluation of the upper bound of the TVD between $K_{0:n}$ and $K_{0:\infty}$ for $\theta \geq 1$ as follows:

THEOREM 2.5. *For $\theta \geq 1$,*

$$d_{TV}(K_{0:n}, K_{0:\infty}) \leq \min\left\{1, \frac{\theta^2}{[\theta]}\Big(\frac{1}{n+1} + \cdots + \frac{1}{n+[\theta]}\Big) + \frac{\theta}{n+\theta}\right\}. \tag{16}$$

According to Theorems 2.4 and 2.5, the rate of convergence $K_{0:n}$ to $K_{0:\infty}$ by the TVD is the following:

COROLLARY 2.6.

$$d_{TV}(K_{0:n}, K_{0:\infty}) = O\Big(\frac{1}{n}\Big).$$

$K_{0:n}$ converges to $K_{0:\infty}$ in TVD as $n \to \infty$. Therefore, the p.f. of $K_{0:n}$ converges pointwise to the p.f. of $K_{0:\infty}$, that is,

$$\lim_{n\to\infty} P(K_{0:n} = k) = P(K_{0:\infty} = k) \quad (k = 0, 1, 2, \dots).$$

We illustrate $K_{0:\infty}$ and $K_{0:n}$ by their p.f.'s $P(K_{0:\infty} = k)$ and $P(K_{0:n} = k)$, respectively. They are plotted for $\theta = 1/2$, $\theta = 2$ and $\theta = 5$, in Figs. 1−3, respectively. For $\theta = 1/2$, the p.f. of $K_{0:n}$ $(n = 5)$ is indicated by a dotted line. For $\theta = 2$, the p.f. of $K_{0:n}$ $(n = 10$ and $n = 30)$ is indicated by a dashed or dotted line, respectively. For $\theta = 5$, the p.f. of $K_{0:n}$ $(n = 10$ and $n = 50)$ is indicated by a dashed or dotted line, respectively. The figures showing these p.f.'s are based on histograms made by the random numbers of the R, using equation (2). The p.f.s of $K_{0:\infty}$ are given by Appendix (A.13) and Appendix (A.14) and indicated by solid lines.

Using the values to draw Figs. 1−3, the approximate values of $d_{TV}(K_{0:n}, K_{0:\infty})$ are obtained. These values and the corresponding upper bounds are listed in the second and fourth rows of Table 1, respectively. As upper bounds of TVD, we used the right-hand side of equation (13) for $\theta = 0.5$ and (16) for $\theta = 2, 5$. The upper bounds of equations

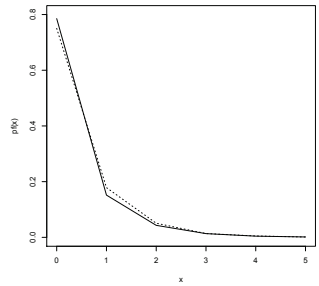

Fig. 1: $\theta = 0.5$; $K_{0:n}$, $n = 5$ (dot)

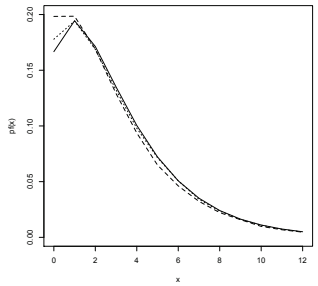

Fig. 2: $\theta = 2$; $K_{0:n}$, $n = 10$ (dash), 30 (dot)

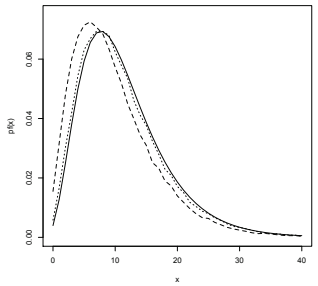

Fig. 3: $\theta = 5$; $K_{0:n}$, $n = 10$ (dash), 50 (dot)

(13) and (16) are considerably larger than the approximate TVDs. To compare with these values, we considered

$$d_{TV}(\widehat{G}_{n:n}, \widetilde{G}_{n:\infty}) = \frac{1}{2}\sum_{j=0}^{\infty} \big|P(\widehat{G}_{n:n} = j) - P(\widetilde{G}_{n:\infty} = j)\big|,$$

where $P(\widehat{G}_{n:n} = j) = p_n(1-p_n)^j$ ($p_n = n/(n+\theta)$ and $\widetilde{G}_{n:\infty}$ have the mixed Poisson distribution $Po(\theta\Lambda_{n:\infty})$. The p.f. of $\widetilde{G}_{n:\infty}$ is indicated in Appendix (A.11) for $\theta > 0$. If θ is a positive integer, then the p.f. of $\widetilde{G}_{n:\infty}$ is indicated by Appendix (A.12).

For example, in case of $\theta = 1$, $P(\widetilde{G}_{n:\infty} = j) = n/(n+1)^{j+1}$ is equal to $P(\widehat{G}_{n:n} = j)$. Therefore, we have

$$d_{TV}(K_{0:n}, K_{0:\infty}) = d_{TV}(\widehat{G}_{n:n}, \widetilde{G}_{n:\infty}) = 0, \tag{17}$$

and

$$K_{0:n} \stackrel{d}{=} K_{0:\infty} \sim Ge(1/2).$$

In case of $\theta = 2$, $P(\widehat{G}_{n:n} = j) = n2^j/(n+2)^{j+1}$ and

$$P(\widetilde{G}_{n:\infty} = k) = \frac{n(n+1)}{2}\Big\{\Big(\frac{2}{n+2}\Big)^{k+1} - \Big(\frac{2}{n+3}\Big)^{k+1}\Big\}. \tag{18}$$

For the third law of the table, we used $P(\widehat{G}_{n:n} = j) = n\theta^j/(n+\theta)^{j+1}$, and Appendix (A.11) with $\theta = 0.5$, equation (18) and Appendix (A.12) with $\theta = 5$ for $\widetilde{G}_{n:\infty}$.

Table 1 . Approximate $d_{TV}(K_{0:n}, K_{0:\infty})$, $d_{TV}(\widehat{G}_{n:n}, \widetilde{G}_{n:\infty})$ and the upper bounds.

	$\theta = 0.5, n = 5$	$\theta = 2,\ n = 10\ (30)$	$\theta = 5, n = 10\ (50)$
Approximate TVD	0.0354	0.0350 (0.0112)	0.1136 (0.0256)
$d_{TV}(\widehat{G}_{n:n}, \widetilde{G}_{n:\infty})$	0.0422	0.1282(0.0568)	0.4942(0.2772)
Upper bound for TVD	0.1409	0.5152 (0.1895)	1 (0.5629)

For $\theta = 2$ and $j = 0$, because of $P(\widehat{G}_{n:n} = 0) - P(\widetilde{G}_{n:\infty} = 0) = \big(2/(n+2)\big)\cdot\big(1-3/(n+3)\big)$ we have

$$d_{TV}(\widehat{G}_{n:n}, \widetilde{G}_{n:\infty}) = \frac{1}{2}\sum_{j=0}^{\infty} \big|P(\widehat{G}_{n:n} = j) - P(\widetilde{G}_{n:\infty} = j)\big| > \frac{1}{n+2}\Big(1 - \frac{3}{n+3}\Big). \tag{19}$$

Thus, for $\theta = 2$ and $n > 4$, equations (15) and (19), we have

$$\frac{1}{n+2}\Big(1 - \frac{3}{n+3}\Big) < d_{TV}(\widehat{G}_{n:n}, \widetilde{G}_{n:\infty}) \le \frac{2}{n+1} + \frac{4}{n+2},$$

which means that $d_{TV}(\widehat{G}_{n:n}, \widetilde{G}_{n:\infty}) = O(1/n)$ $(\theta = 2)$. Therefore, we cannot get the upper bound for $d_{TV}(K_{0:n}, K_{0:\infty})$ strictly less than the order $O(1/n)$ for all $\theta > 0$, by using $d_{TV}(\widehat{G}_{n:n}, \widetilde{G}_{n:\infty})$ and equation (9). The lower bound for $d_{TV}(K_{0:n}, K_{0:\infty})$ is zero, resulting from $\theta = 1$ as shown in equation (17). To obtain a more accurate upper bound for $d_{TV}(K_{0:n}, K_{0:\infty})$.

3. Appendix

3.1. On Lemma 1.1

We explain Lemma 1.1 following Pitman and Yakubovich (2017, 2019). For $\mathbf{P} = (P_1, P_2, \dots)$ on $\mathbf{N}$ having a GEM distribution with parameter $\theta(> 0)$, its distribution can be expressed as $F_0 = 0$ and $F_j = \sum_{i=1}^{j} P_i = 1 - \prod_{i=1}^{j}(1-\xi_i)$ $(j = 1, 2, \dots)$ and put

$$N_F(a, b] = \sum_{i=1}^{\infty} I(a < F_i \le b) \ \ (0 \le a < b < 1),$$

where $I(A)$ is the indicator function of an event A.

For the sample $U_1, \dots, U_n$ of size n from the uniform distribution $U(0, 1)$, we use $X_i = N_F(0, U_i] + 1 \quad (i = 1, \dots, n)$. Then, we get

$$P(X_i = j \mid \mathbf{P}) = P_j \quad (i = 1, \dots, n \ ; \ j = 1, 2, \dots).$$

That is, $X_1, \dots, X_n$ can be regarded as the sample from random frequencies $\mathbf{P} = (P_1, P_2, \dots)$ on $\mathbf{N}$. Let

$$S_0 = 0 \quad \text{and} \quad S_j = -\log(1 - F_j) = \sum_{i=1}^{j} -\log(1 - \xi_i) \ (j = 1, 2, \dots).$$

The point process associated with S_j $(j = 0, 1, 2, \dots)$ can be expressed as follows:

$$N_S(s, t] = \sum_{j=1}^{\infty} I(s < S_j \le t) \quad (0 \le s < t < \infty).$$

We have $0 < S_1 < S_2 < \cdots \uparrow \infty$ a.s. Because ξ_i $(i = 1, 2, \dots)$ are independent and identically distributed with $Beta(1, \theta)$, $-\log(1 - \xi_i)$ $(i = 1, 2, \dots)$ are independent and identically distributed, with the exponential distribution $Exp(\theta)$. Therefore, the point process N_S on $(0, \infty)$ has a stationary Poisson process with intensity θ (Feller (1966, pp.10−11) and Billingsley (1995, pp.297−299)).

Let $U_{1:n}, \dots, U_{n:n}$ be the order statistics of the sample $U_1, \dots, U_n$ from $U(0, 1)$. Then, $\varepsilon_{i:n} = -\log(1 - U_{i:n})$ $(i = 1, 2, \dots, n)$ are the order statistics based on the standard exponential distribution $Exp(1)$. Using the relations $X_i = N_F(0, U_i] + 1$ and $N_S(s, t] = N_F(1 - e^{-s}, 1 - e^{-t}] \quad (0 \le s < t < \infty)$, we have

$$X_{i:n} - 1 = N_F(0, U_{i:n}] = N_F(0, 1 - e^{-\varepsilon_{i:n}}] = N_S(0, \varepsilon_{i:n}].$$

As the Poisson process N_S is stationary, we have

$$\widehat{G}_{i:n} = X_{n-i+1:n} - X_{n-i:n} = N_S(\varepsilon_{n-i:n}, \varepsilon_{n-i+1:n}] \stackrel{d}{=} N_S(0, \varepsilon_{n-i+1:n} - \varepsilon_{n-i:n}].$$

The differences $\varepsilon_{n-i+1:n} - \varepsilon_{n-i:n}$ $(i = 1, 2, \ldots, n)$ in order statistics based on $Exp(1)$ are independent and have the exponential distribution $Exp(1/i)$ (Arnold et al. (1972, pp.72–73)). Therefore, $\widehat{G}_{i:n}$ $(i = 1, 2, \ldots, n)$ are independent and $\widehat{G}_{i:n} \stackrel{d}{=} N_S(0, \varepsilon/i)$, where ε has $Exp(1)$. As the intensity of the Poisson process N_S is θ, $\widehat{G}_{i:n}$ has the mixed Poisson distribution $Po\big(\theta\varepsilon/i\big)$, which is equivalent to the geometric distribution $Ge(i/(i+\theta))$.

3.2. Expression of $K_{0;n}$

The terms $\widehat{G}_{n:n}$ and $(\widehat{G}_{i:n} - 1)_+$ of equation (2) are explained as follows:

(i) $\widehat{G}_{n:n}$; In case of $X_{1:n} = 1$, there is no positive integer less than $X_{1:n}$. Thus, $\widehat{G}_{n:n} = X_{1:n} - 1 = 0$. In case of $X_{1:n} \geq 2$, there are $\widehat{G}_{n:n} = X_{1:n} - 1$ positive integers less than $X_{1:n}$.

(ii) $(\widehat{G}_{i:n} - 1)_+$ $(i = 1, \ldots, n-1)$; in case of $\widehat{G}_{i:n} = X_{n-i+1:n} - X_{n-i:n} = 0$ or 1, there is no positive integer greater than $X_{n-i:n}$ and less than $X_{n-i+1:n}$. In case of $\widehat{G}_{i:n} \geq 2$, there are $\widehat{G}_{i:n} - 1$ positive integers greater than $X_{n-i:n}$ and less than $X_{n-i+1:n}$. Therefore, $(\widehat{G}_{i:n} - 1)_+$ denotes the numbers of unobserved positive integers greater than $X_{n-i:n}$ and less than $X_{n-i+1:n}$.

By **(i)** and **(ii)**, we have the expression of $K_{0:n}$ resulting from equation (2).

3.3. Distribution of $(\widehat{G}_{j:n} - 1)_+$

Per the definition of $(\widehat{G}_{j:n} - 1)_+$, we have

$$(\widehat{G}_{j:n} - 1)_+ = \begin{cases} 0 & (\widehat{G}_{j:n} = 0) \\ \widehat{G}_{j:n} - 1 & (\widehat{G}_{j:n} \geq 1) \end{cases}.$$

Since $\widehat{G}_{j:n}$ has the geometric distribution $Ge(j/(j+\theta))$ and $P(\widehat{G}_{j:n} \geq 1) = \theta/(j+\theta)$, for $k = 0, 1, 2, \ldots,$

$$P((\widehat{G}_{j:n} - 1)_+ = k \mid \widehat{G}_{j:n} \geq 1) = P(\widehat{G}_{j:n} = k+1 \mid \widehat{G}_{j:n} \geq 1) = \left(\frac{j}{j+\theta}\right)\left(\frac{\theta}{j+\theta}\right)^k.$$

The right-hand side is the p.f of the geometric distribution $Ge\big(j/(j+\theta)\big)$, which is equivalent to the mixed Poisson distribution $Po(\theta\varepsilon_j/j)$. As the Bernoulli variable B_{q_j} has $Ber(q_j)$, $q_j = P(\widehat{G}_{j:n} \geq 1)$ and is independent of ε_j, the distribution of $(\widehat{G}_{j:n} - 1)_+$ can be indicated as the mixed Poisson distribution $Po(\theta B_{q_j}\varepsilon_j/j)$, which is the mixed Poisson distribution $Po(\theta\varepsilon_j/j)$ for $B_{q_j} = 1$ and degenerate at 0 for $B_{q_j} = 0$.

3.4. Sum of independent mixed Poisson variables

We consider the distribution of the sum of independent mixed Poisson distributions.

PROPOSITION A.1. *Let Y_1 and Y_2 be independent random variables with the mixed Poisson distributions $Po(\Lambda_1^*)$ and $Po(\Lambda_1^*)$, respectively. Let the mixing parameters Λ_1^* and Λ_2^* be independent. Then, Y_1+Y_2 has the mixed Poisson distribution $Po(\Lambda_1^*+\Lambda_2^*)$, that is,*

$$Y_1+Y_2 \sim Po(\Lambda_1^*+\Lambda_2^*).$$

Proof. Given the independence of Y_1 and Y_2, the probability generating function (p.g.f.) of Y_1+Y_2 is

$$\begin{aligned} E[z^{Y_1+Y_2}] &= E[z^{Y_1}]\cdot E[z^{Y_2}] = E\big[E[z^{Y_1}\mid \Lambda_1^*]\big]\cdot E\big[E[z^{Y_2}\mid \Lambda_2^*]\big] \\ &= E[e^{\Lambda_1^*(z-1)}]\cdot E[e^{\Lambda_2^*(z-1)}] = E[e^{(\Lambda_1^*+\Lambda_2^*)(z-1)}], \end{aligned}$$

where the last equation holds by the independence of Λ_1^* and Λ_2^*. As the last expression is the p.g.f. of $Po(\Lambda_1^*+\Lambda_2^*)$, we get the desired result. □

In general, we have the following:

COROLLARY A.2. *Let $Y_1, Y_2, \dots, Y_r$ be independent random variables with the mixed Poisson distributions $Po(\Lambda_1^*), Po(\Lambda_2^*), \dots, Po(\Lambda_r^*)$, for $r \geq 2$. Let the mixing parameters $\Lambda_1^*, \Lambda_2^*, \dots, \Lambda_r^*$ be independent. Then,*

$$Y_1+Y_2+\cdots+Y_r \sim Po(\Lambda_1^*+\Lambda_2^*+\cdots+\Lambda_r^*) \quad (r=2,3,\dots).$$

3.5. Distribution of $\Lambda_{n:\infty}$

Let $B_{q_j^*}(j=0,1,2,\dots)$ be independent random variables having the Bernoulli distribution $Ber(q_j^*)$ ($q_j^* = b/(a+b+j)$; $a \geqq 0$, $b>0$) and ε_j ($j=0,1,2,\dots$) be mutually independent random variables having the exponential distribution $Exp(1)$. In addition, let $\{B_{q_j^*}\}$ and $\{\varepsilon_j\}$ be mutually independent. For $c,d>0$, $\beta_{c,d}$ denotes a random variable having the beta distribution $Beta(c,d)$.

PROPOSITION A.3. *For $n=1,2,\dots$ and $a \geqq 0$, $b>0$,*

$$\exp\Big(-\sum_{j=n}^{\infty} B_{q_j^*}\frac{\varepsilon_j}{a+j}\Big) \sim Beta(a+n,b). \tag{A.1}$$

Therefore,

$$\sum_{j=n}^{\infty} B_{q_j^*}\frac{\varepsilon_j}{a+j} \stackrel{d}{=} -\log\beta_{a+n,b}. \tag{A.2}$$

Proof. First, we write $\sum_{j=n}^{\infty} B_{q_j^*}\varepsilon_j/(a+j) < \infty$ a.s., based on

$$0 \leq E\Big[\sum_{j=n}^{\infty} B_{q_j^*}\frac{\varepsilon_j}{a+j}\Big] \leq \sum_{j=n}^{\infty}\frac{b}{a+b+j}\cdot\frac{1}{a+j} < \frac{\pi^2}{6}b\,.$$

Therefore, we have

$$0 \le \exp\Big(-\sum_{j=n}^{\infty} B_{q_j^*}\frac{\varepsilon_j}{a+j}\Big) \le \cdots \le \exp\Big(-\sum_{j=n}^{n+2} B_{q_j^*}\frac{\varepsilon_j}{a+j}\Big)$$
$$\le \exp\Big(-\sum_{j=n}^{n+1} B_{q_j^*}\frac{\varepsilon_j}{a+j}\Big) \le 1,\ a.s. \qquad (A.3)$$

We consider the rth moment of $\exp\{-\sum_{j=n}^{m} B_{q_j^*}\varepsilon_j/(a+j)\}$ $(m > n)$ about the origin as follows:

$$E\left\{\left[\exp\Big(-\sum_{j=n}^{m} B_{q_j^*}\frac{\varepsilon_j}{a+j}\Big)\right]^r\right\} = E\left[\exp\Big(-\sum_{j=n}^{m} r\cdot B_{q_j^*}\frac{\varepsilon_j}{a+j}\Big)\right]\ (r=1,2,\dots). \quad (A.4)$$

Based on the independence of $\{B_{q_j^*}\}$, $\{\varepsilon_j\}$, and $Ee^{-s\varepsilon_j} = 1/(1+s)$ $(s > -1)$, Appendix (A.4) becomes

$$\prod_{j=n}^{m} E\Big[\exp\Big(-\frac{r}{a+j}B_{q_j^*}\times\varepsilon_j\Big)\Big]$$
$$= \prod_{j=n}^{m} E\left[\frac{1}{1+\frac{r}{a+j}B_{q_j^*}}\right] = \prod_{j=n}^{m}\frac{a+j}{a+b+j}\cdot\frac{a+b+r+j}{a+r+j}$$
$$= \frac{\Gamma(a+b+n)\Gamma(a+r+n)}{\Gamma(a+n)\Gamma(a+b+r+n)}\times\left[\frac{\Gamma(a+m+1)}{\Gamma(a+b+m+1)}\cdot\frac{\Gamma(a+b+r+m+1)}{\Gamma(a+r+m+1)}\right]. \qquad (A.5)$$

As $m \to \infty$, Appendix (A.5) converges to $B(a+n+r,b)/B(a+n,b)$, because the brackets of Appendix (A.5) converge to 1 . Based on Appendix (A.3) and the Monotone convergence theorem, as $m \to \infty$ Appendix (A.4) converges. Thus, we have the following:

$$E\left\{\left[\exp\Big(-\sum_{j=n}^{\infty} B_{q_j^*}\frac{\varepsilon_j}{a+j}\Big)\right]^r\right\} = \frac{B(a+n+r,b)}{B(a+n,b)} \quad (r=1,2,\dots). \qquad (A.6)$$

The right-hand side is the rth moment of the beta distribution $Beta(a+n,b)$ about the origin. Because the beta distribution is determined uniquely by the moments, Appendix (A.6) implies that $\exp\{-\sum_{j=n}^{\infty} B_{q_j^*}\varepsilon_j/(a+j)\}$ has the beta distribution $Beta(a+n,b)$. Thus we obtain Appendix (A.1). □

By applying $a = 0$ and $b = \theta$ to Proposition A.3, we have the following relation:

COROLLARY A.4. *For* $n = 1, 2, \dots,$

$$\exp\Big(-\sum_{j=n}^{\infty} B_{q_j}\frac{\varepsilon_j}{j}\Big) \sim Beta(n,\theta). \qquad (A.7)$$

Therefore,

$$\Lambda_{n:\infty} = \sum_{j=n}^{\infty} B_{q_j}\frac{\varepsilon_j}{j} \overset{d}{=} -\log\beta_{n,\theta}. \qquad (A.8)$$

Especially,

$$\Lambda_{1:\infty} = \sum_{j=1}^{\infty} B_{q_j} \frac{\varepsilon_j}{j} \stackrel{d}{=} -\log \xi. \tag{A.9}$$

Pitman and Yakubovich (2017, p.20) states that the relation Appendix (A.9) can be proven by computing the Laplace transform. Proposition 7.1. indicated the relation Appendix (A.2) in case of $a > 0$ and $n = 0$, based on the additivity of Lévy measures.

3.6. Probability function of $\widetilde{G}_{n:\infty}$

By Appendix (A.8) and $\beta_{n,\theta} \sim Beta(n, \theta)$, the density function of $\Lambda_{n:\infty}$ is given by

$$f_{\Lambda_{n:\infty}}(t) = \frac{1}{B(n,\theta)} e^{-nt}(1 - e^{-t})^{\theta - 1} \quad (t > 0).$$

Therefore, for $k = 0, 1, 2, \ldots$, the p.f. of $\widetilde{G}_{n:\infty}$ having the mixed Poisson distribution $Po(\theta \Lambda_{n:\infty})$ is

$$\begin{aligned} P(\widetilde{G}_{n:\infty} = k) &= \frac{\theta^k}{k!} \cdot \frac{1}{B(n,\theta)} \int_0^{\infty} t^k (1 - e^{-t})^{\theta - 1} e^{-(\theta + n)t} dt \\ &= \frac{\theta^k}{k!} \cdot \frac{1}{B(n,\theta)} \int_0^{\infty} \sum_{j=0}^{\infty} (-1)^j \binom{\theta - 1}{j} t^k e^{-(\theta + j + n)t} dt \\ &= \frac{\theta^k}{k!} \cdot \frac{1}{B(n,\theta)} \sum_{j=0}^{\infty} (-1)^j \binom{\theta - 1}{j} \frac{1}{(\theta + j + n)^{k+1}} \int_0^{\infty} \tau^k e^{-\tau} d\tau. \end{aligned} \tag{A.10}$$

As the integral of Appendix (A.10) is equal to $k!$, the p.f. of $\widetilde{G}_{n:\infty}$ is written as follows:

$$P(\widetilde{G}_{n:\infty} = k) = \frac{1}{\theta B(n,\theta)} \sum_{j=0}^{\infty} (-1)^j \binom{\theta - 1}{j} \left(\frac{\theta}{\theta + j + n} \right)^{k+1} \quad (k = 0, 1, 2, \ldots). \tag{A.11}$$

If θ is a positive integer l, then

$$P(\widetilde{G}_{n:\infty} = k) = \frac{1}{l B(n,l)} \sum_{j=0}^{l-1} (-1)^j \binom{l - 1}{j} \left(\frac{l}{l + j + n} \right)^{k+1} \quad (k = 0, 1, 2, \ldots). \tag{A.12}$$

According to $K_{0:\infty} \sim Po(-\theta \log \xi)$ and Appendix (A.9), we have the equivalent relation $\widetilde{G}_{1:\infty} \stackrel{d}{=} K_{0:\infty}$ in distribution. Therefore, based on Appendix (A.11) and (A.12) with $n = 1$, for $\theta > 0$ we have the following:

$$P(K_{0:\infty} = k) = \sum_{j=0}^{\infty} (-1)^j \binom{\theta - 1}{j} \left(\frac{\theta}{\theta + j + 1} \right)^{k+1} \quad (k = 0, 1, 2, \ldots). \tag{A.13}$$

If θ is the positive integer l, then

$$P(K_{0:\infty} = k) = \sum_{j=0}^{l-1} (-1)^j \binom{l - 1}{j} \left(\frac{l}{l + j + 1} \right)^{k+1} \quad (k = 0, 1, 2, \ldots). \tag{A.14}$$

Acknowledgement

The author is grateful to the referee for his/her careful reading and useful comments.

References

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics* **2** (6), 1152–1174.

Arnold, B. C., Balakrishnan, A. and Nagaraja, H. N. (1992). *A first course in order statistics*. John Wiley & Sons, New York.

Billingsley, P. (1995). *Probability and measure*. John Wiley & Sons, New York.

Engen, S. (1975). A note on the geometric series as species frequency model, *Biometrika* **62**, 697–699.

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

Ewens, W.J. (1990). Population genetics theory-the past and the future, in: Mathematical and statistical developments of evolutionary theory (ed. by Lessard, S.), 177–227.

Feller, W. (1966). *An introduction to probability theory and its applications* Vol. 2, John Wiley & Sons, New York.

Freedman, D. (1974). The Poisson approximation for dependent events, *Annals of Probability* **2**, 256–269.

Gnedin, A., Iksanov, A., Negadajlov, P. and Rösler. U. (2009). The Bernoulli sieve revisited, *The Annals of Applied Probability*, **19** (4), 1634–1655.

Gnedin, A., Iksanov. A., and Marynych, A. (2010). The Bernoulli sieve: overview, arXiv:1005.5705 [math.PR].

Grandell. J. (1997). *Mixed Poisson processes*, Chapman & Hall, London.

Johnson, N. L., Kotz, S. and Balakrishnan (1997). *Discrete multivariate distributions*, Wiley & Sons, New York.

Johnson, N. L., Kemp, A. W. and Kotz, S. (2005). *Univariate discrete distributions*, 3rd Ed., Wiley & Sons, New Jersey.

McCloskey. J. W. (1965). A model for the distribution of individuals by species in an environment, Ph. D. thesis, Michigan State Univ.

Pitman, J. and Yakubovich, Y. (2017). Extremes and gaps in sampling from a GEM random discrete distribution, *Electronic Journal of Probability*, **22**, no. **44**, 1–26.

Pitman, J. and Yakubovich, Y. (2019). Gaps and and interleaving of point processes in sampling from a residual allocation model, *Bernoulli*, **25** (4B), 3623–3651.

Yamato, H. (2020). *Statistics based on Dirichlet processes and related topics* (JSS Research Series in Statistics), Springer, Singapore.

Received: April 1, 2021
Revised: August 16, 2021; October 17, 2022
Accept: October 19, 2022