Ann Inst Stat Math (2011) 63:745–768
DOI 10.1007/s10463-010-0276-7

Limiting size index distributions for ball-bin models with Zipf-type frequencies

Satoshi Chida · Naoto Miyoshi

Received: 30 January 2008 / Revised: 3 April 2009 / Published online: 5 February 2010

Abstract We consider a random ball-bin model where balls are thrown randomly and sequentially into a set of bins. The frequency of choices of bins follows the Zipf-type (power-law) distribution; that is, the probability with which a ball enters the ith most popular bin is asymptotically proportional to $1/i^{\alpha}$, $\alpha > 0$. In this model, we derive the limiting size index distributions to which the empirical distributions of size indices converge almost surely, where the size index of degree k at time t represents the number of bins containing exactly k balls at t. While earlier studies have only treated the case where the power α of the Zipf-type distribution is greater than unity, we here consider the case of $\alpha \leq 1$ as well as $\alpha > 1$. We first investigate the limiting size index distributions for the independent throw models and then extend the derived results to a case where bins are chosen dependently. Simulation experiments demonstrate not only that our analysis is valid but also that the derived limiting distributions well approximate the empirical size index distributions in a relatively short period.

Keywords Limiting distributions · Random ball-bin occupancy models · Size indices · Zipf-type distribution

1 Introduction

We consider a random ball-bin occupancy model where balls are thrown randomly and sequentially into a set of bins. One of the most important criteria characterizing the features of such models is the distribution of *size indices*, which are also called *frequency*

S. Chida · N. Miyoshi (✉)
Department of Mathematical and Computing Sciences, Tokyo Institute of Technology,
2-12-1-W8-52 Ookayama, Tokyo 152-8552, Japan
e-mail: miyoshi@is.titech.ac.jp

spectra, *spectral statistics* or *frequencies of frequency* (see, e.g., Sibuya (1993) and Baayen (2001)). Let m ($\in \overline{\mathbb{N}} = \mathbb{N} \cup \{+\infty\}$) denote the number of bins and let $N_i^{(m)}(t)$, $i = 1, \dots, m, t \geq 0$, denote the number of balls in bin i at time t, where we assume that $N_i^{(m)}(0) = 0, i = 1, \dots, m$; that is, all bins are initially empty. The size index $S_k^{(m)}(t)$ of degree k ($= 1, 2, \dots$) at time t is then defined as $S_k^{(m)}(t) = \sum_{i=1}^{m} 1_{\{N_i^{(m)}(t)=k\}}$, where 1_E denotes the indicator function for event E; that is, $S_k^{(m)}(t)$ denotes the number of bins containing exactly k balls at t. The size indices are indeed important for applications, for example, to efficient coding or data compression of computerized texts, where $S_k^{(m)}(t)$ represents the number of distinct letters/words appearing exactly k times in a given reading of a text. In this paper, we investigate the limiting distributions of size indices with respect to an increase of time (and an increase of the number of bins when $m < \infty$); that is, we derive the almost sure limit of $S_k^{(\infty)}(t)/\overline{S}_1^{(\infty)}(t)$, $k = 1, 2, \dots$, as $t \to \infty$ when $m = +\infty$ and that of $S_k^{(m)}(t(m))/m$, $k = 0, 1, 2, \dots$, as $m \to \infty$ when $m < \infty$, where $\overline{S}_k^{(m)}(t) = \sum_{l \geq k} S_l^{(m)}(t)$, $S_0^{(m)}(t) = m - \overline{S}_1^{(m)}(t)$ and $t(m)$ is an appropriate function of m satisfying $t(m) \to \infty$ as $m \to \infty$.

Studies concerning the asymptotics of size indices date back from the 1960s to the 1970s. Karlin (1967) considered the case where the number of bins was infinite and the choices of bins were i.i.d. for both the discrete-time model; that is, a ball is thrown at every time unit, and the Poisson embedded model; that is, balls are thrown according to a homogeneous Poisson process. Among many other results, he derived the asymptotics of $\mathrm{E}S_k^{(\infty)}(t)$, $k = 1, 2, \dots$, as $t \to \infty$ and also showed that $\overline{S}_k^{(\infty)}(t)/\mathrm{E}\overline{S}_k^{(\infty)}(t) \to 1$ almost surely as $t \to \infty$, where E denotes the expectation. While it is easy to obtain the almost sure limit of the empirical size index distribution $S_k^{(\infty)}(t)/\overline{S}_1^{(\infty)}(t)$, $k = 1, 2, \dots$, as $t \to \infty$ from Karlin's results, Rouault (1978) further extended this to the case where the choices of bins are Markov dependent assuming that the frequency distribution follows a generalized Zipf-type law; that is, a distribution with a regularly varying tail. A recent survey of infinite bin models is presented in Gnedin et al. (2007), which includes the case that the probability distribution of the choices of bins is randomly but independently varying, and a survey of models with finite bins is found in Ivanov et al. (1985).

Past studies have extensively developed frequency models with a Zipf-like power-law in various fields (see, e.g., Baayen (2001) and references therein for lexical models). We therefore employ a Zipf-type distribution as the frequency model; that is, the probability with which a ball enters bin i is asymptotically proportional to $1/i^{\alpha}$ with $\alpha > 0$ for large i and m. Note here that, if the number m of bins is infinite, then the power α of the Zipf-type distribution must be greater than unity. In this paper, we extend the results in Karlin (1967), Rouault (1978) and Gnedin et al. (2007), that are mentioned above, in two directions. In the first, we consider the case of $\alpha \leq 1$ and $m < \infty$, where the number m of bins increases together with time. We illustrate that, by determining an appropriate function $t(m)$, $m \in \mathbb{N}$, we can obtain the almost sure limit of the empirical size index distribution $S_k^{(m)}(t(m))/m$, $k = 0, 1, 2, \dots$, as $m \to \infty$. We also discuss the relation of the results with Khmaladze's definitions of a large number of rare events (LNRE) (see Khmaladze (1987) and also Baayen (2001)). The second extension is where the choices of bins are generally dependent while their marginal distributions are common Zipf-type ones, and we demonstrate that the results derived for the independent throw model are still valid for the dependent throw model under some additional conditions. The model we consider here is the Poisson embedded one; that is, balls are thrown according to a homogeneous Poisson process with intensity 1, and at each throw, the ball enters one of the bins according to the given Zipf-type probability.

The organization of the paper is as follows. The next section takes into consideration the independent throw model, where we first describe the model and review the existing results for the case of $\alpha > 1$ and $m = +\infty$. We then consider the cases of $\alpha < 1$ and $\alpha = 1$ with $m < +\infty$ and derive the limiting size index distributions in the form of $\lim_{m\to\infty} S_k^{(m)}(t(m))/m$, $k = 0, 1, 2, \ldots$, for some appropriately chosen $t(m)$, $m \in \mathbb{N}$. In Sect. 3, we extend the model such as the choices of bins are dependent in some general sense and verify that all results in Sect. 2 are still valid under some additional conditions. The derived results are validated through simulation experiments in Sect. 4, where we also observe that the limiting distributions well approximate the empirical size index distributions in a relatively short period.

2 Independent throw model

Throughout this and the next sections, we have assumed that all random elements are defined on a common probability space $(\Omega, \mathcal{F}, \mathrm{P})$. In the analysis, we use the following standard notations. For any two real functions $f(x)$ and $g(x)$, $x \in \mathbb{R}$, $f(x) \sim g(x)$ as $x \to a$ stands for $\lim_{x\to a} f(x)/g(x) = 1$, where a is possibly $+\infty$. Also, we write $f(x) = \Theta(g(x))$ as $x \to a$ for $0 < \liminf_{x\to a} |f(x)/g(x)| \le \limsup_{x\to a} |f(x)/g(x)| < \infty$ and $f(x) = o(g(x))$ as $x \to a$ for $\lim_{x\to a} f(x)/g(x) = 0$.

2.1 Model description

Let $\{N(t)\}_{t\ge 0}$ denote a homogeneous Poisson process with intensity 1, where $N(t)$, $t \ge 0$, represents the number of points of the Poisson process during $(0, t]$. At each point of $\{N(t)\}_{t\ge 0}$, a ball is thrown randomly into one of the bins, where the set of bins is $\{1, \ldots, m\}$ when $m < +\infty$, or $\{1, 2, \ldots\}$ when $m = +\infty$. We here assume that all bins are initially empty. Let B_n, $n \in \mathbb{N}$, denote the random variable representing the bin that the nth ball enters. We assume that B_n, $n \in \mathbb{N}$, are mutually independent and are also independent of the Poisson process $\{N(t)\}_{t\ge 0}$. The probability with which a ball enters bin i $(= 1, \ldots, m)$ is denoted by $p_i^{(m)} = \mathrm{P}(B_1 = i)$, where $p_i^{(m)} \ge 0$, $i = 1, \ldots, m$, and $p_1^{(m)} + \cdots + p_m^{(m)} = 1$. Let $N_i^{(m)}(t)$, $i = 1, \ldots, m$, denote the number of balls in bin i at time t, so that $\sum_{i=1}^m N_i^{(m)}(t) = N(t)$, $t \ge 0$. By the fundamental property of Poisson processes, $N_i^{(m)}(t)$, $i = 1, \ldots, m$, are mutually independent and also follow Poisson distributions with mean $p_i^{(m)} t$; that is,

$$\mathrm{P}\left(N_1^{(m)}(t) = k_1, \ldots, N_m^{(m)}(t) = k_m\right) = \prod_{i=1}^{m} \frac{(p_i^{(m)}\, t)^{k_i}}{k_i!}\, e^{-p_i^{(m)} t}, \quad t \ge 0,\ k_1, \ldots, k_m \in \mathbb{Z}_+.$$

We assume that the frequency distribution $\boldsymbol{p}^{(m)} = (p_1^{(m)}, \ldots, p_m^{(m)})$ is Zipf-type; that is, $p_i^{(m)}$ is asymptotically proportional to $1/i^{\alpha}$, $\alpha > 0$, for large i and m. Within this setting, we will next investigate the limiting distributions of size indices $S_k^{(m)}(t) = \sum_{i=1}^{m} 1_{\{N_i^{(m)}(t)=k\}}$, $k = 1, 2, \ldots$, for the cases of $\alpha > 1$, $\alpha < 1$ and $\alpha = 1$ separately.

2.2 Case of $\alpha > 1$

In this subsection, we review the results for the independent throw model with Zipf-type frequency in the case where the number m of bins is infinite and the power α of the Zipf-type distribution is greater than unity. The results are obtained directly from Karlin (1967) and Rouault (1978) (see also Gnedin et al. (2007)). We here suppress the superscript "(∞)" and write, for example, p_i, $N_i(t)$ for $p_i^{(\infty)}$, $N_i^{(\infty)}(t)$, and so on. In this case, the Zipf-type frequency distribution $\boldsymbol{p} = (p_1, p_2, \ldots)$ is as follows.

Assumption 1 $p_i \sim c/i^{\alpha}$ as $i \to \infty$ with $\alpha > 1$ and $c > 0$. That is, for any $\epsilon > 0$, there exists an integer $i_{\epsilon} > 0$ such that, for all $i \ge i_{\epsilon}$, inequality $(1-\epsilon)\, c/i^{\alpha} \le p_i \le (1+\epsilon)\, c/i^{\alpha}$ holds.

As defined above, let $S_k(t)$, $k = 1, 2, \ldots$, $t \ge 0$, denote the size index of degree k at time t; that is, the number of bins containing exactly k balls at t. Also, let $\overline{S}_k(t) = \sum_{l \ge k} S_l(t)$; that is, the number of bins containing at least k balls at t. Then, Karlin (1967) derived the following result in a more general form.

Proposition 1 *Under* Assumption 1, *we have for* $k = 1, 2, \ldots$,

$$\mathrm{E}\overline{S}_k(t) \sim c^{1/\alpha}\, t^{1/\alpha} \left[\Gamma\left(1 - \frac{1}{\alpha}\right) - \sum_{l=1}^{k-1} \frac{1}{\alpha\, l!}\, \Gamma\left(l - \frac{1}{\alpha}\right)\right] \quad \text{as } t \to \infty, \tag{1}$$

where conventionally $\sum_{l=1}^{0} \cdot = 0$ *and* Γ *denotes Euler's Gamma function; that is,* $\Gamma(x) = \int_0^{\infty} e^{-u}\, u^{x-1}\, \mathrm{d}u$.

Applying $S_k(t) = \overline{S}_k(t) - \overline{S}_{k+1}(t)$ a.s. in Proposition 1, we can readily obtain the limit of $\mathrm{E}S_k(t)/\mathrm{E}\overline{S}_1(t)$, $k = 1, 2, \ldots$, as $t \to \infty$ (which is the same as that given on the right-hand side of (2) below) under Assumption 1. It should be noted that this case does not meet Khmaladze's first definition of a large number of rare events (LNRE) in the sense that $\lim_{t\to\infty} \mathrm{E}S_1(t)/t = 0$ but does the second definition in the sense that $\lim_{t\to\infty} \mathrm{E}\overline{S}_1(t) = \infty$ and $\lim_{t\to\infty} \mathrm{E}S_1(t)/\mathrm{E}\overline{S}_1(t) > 0$ (see Khmaladze (1987) for the original definitions). Karlin (1967) further proved that $S_k(t)/\mathrm{E}S_k(t) \to 1$ a.s. as $t \to \infty$. These results immediately yield the following.

Proposition 2 *Under* Assumption 1, *we have for* $k = 1, 2, \ldots$,

$$\lim_{t\to\infty} \frac{S_k(t)}{\overline{S}_1(t)} = \frac{\Gamma(k - 1/\alpha)}{\alpha\, k!\, \Gamma(1 - 1/\alpha)} = \frac{1}{\alpha\, k} \prod_{l=1}^{k-1} \left(1 - \frac{1}{\alpha\, l}\right) \quad a.s., \tag{2}$$

where $\prod_{l=1}^{0} \cdot = 1$ *conventionally.*

Let $\Psi_k(\alpha), \alpha > 1, k = 1, 2, \ldots$, denote the right-hand side of (2). We can check that $\Psi_k(\alpha), k = 1, 2, \ldots$, gives a proper distribution on $\mathbb{N}$ with $\Psi_k(\alpha) \sim k^{-1-1/\alpha}/[\alpha\,\Gamma(1-1/\alpha)]$ as $k \to \infty$. Rouault (1978) provided the same formula for the discrete-time model and further extended this as it is also valid when the choices of bins are Markov dependent. As the same distribution was derived by Sibuya (1979) independently in another problem, we refer to the distribution given by $\Psi_k(\alpha)$, $k = 1, 2, \ldots$, as the Karlin-Rouault-Sibuya (KRS) distribution. Propositions 1 and 2 are extended in Sect. 3 to the case where the choices of bins are generally dependent.

2.3 Case of $\alpha < 1$

We here consider the case where the power α of the Zipf-type distribution is less than unity and the number m of bins is finite. We derive the limiting size index distribution with respect to an increase of m together with time. The frequency distribution is assumed to satisfy the following.

Assumption 2 For any $\epsilon > 0$, there exists an integer $i_\epsilon > 0$ such that, for all m and i satisfying $m \geq i \geq i_\epsilon$, inequality $(1-\epsilon)\, c/(m^{1-\alpha}\, i^\alpha) \leq p_i^{(m)} \leq (1+\epsilon)\, c/(m^{1-\alpha}\, i^\alpha)$ holds with $\alpha \in (0, 1)$ and $c > 0$.

Note that Assumption 2 represents the Zipf-type distribution with $\alpha < 1$. Indeed, if $p_i^{(m)} = c_m/i^\alpha$ for $i = 1, \ldots, m$ with the normalization constant c_m, we then have $c_m = (\sum_{i=1}^{m} 1/i^\alpha)^{-1} \sim (1-\alpha)/m^{1-\alpha}$ as $m \to \infty$ and Assumption 2 is fulfilled with $c = 1 - \alpha$. Thus, parameter c gives a slight generality to the case of $p_i^{(m)} = c_m/i^\alpha$, $i = 1, \ldots, m$. Under this assumption, we first derive the asymptotic result for the expected size indices, which has attracted independent interest and is indeed also exploited in the extension to the dependent throw model in the next section.

Lemma 1 *Under* Assumption 2, *we have for any constant* $\delta > 0$ *and* $k = 1, 2, \ldots$,

$$\mathrm{E}\overline{S}_k^{(m)}\left(\frac{\delta\, m}{c}\right) \sim m \left[1 - \sum_{l=0}^{k-1} \frac{\delta^{1/\alpha}}{\alpha\, l!}\, \Gamma\left(l - \frac{1}{\alpha},\, \delta\right)\right] \quad as\ m \to \infty, \tag{3}$$

where $\Gamma(x, y)$, $x \in \mathbb{R}$, $y > 0$, *is the incomplete Gamma function; that is,* $\Gamma(x, y) = \int_y^\infty e^{-u}\, u^{x-1}\, \mathrm{d}u$ (*see, e.g.,* Davis (1972)).

Proof Since $\mathrm{E}\overline{S}_k^{(m)}(t) = \sum_{i=1}^m \mathrm{P}(N_i^{(m)}(t) \geq k)$ and $N_i^{(m)}(t)$, $i = 1, \ldots, m$, are Poisson random variables with mean $p_i^{(m)}\,t$, we have

$$\mathrm{E}\overline{S}_k^{(m)}(t) = \sum_{i=1}^m \left[1 - \sum_{l=0}^{k-1} \frac{(p_i^{(m)}\,t)^l}{l!}\, e^{-p_i^{(m)} t} \right].$$

Here, under Assumption 2, for any $\epsilon \in (0, 1)$ and sufficiently large m,

$$\begin{aligned}\mathrm{E}\overline{S}_k^{(m)}(t) &\leq m - \sum_{i=i_\epsilon}^m \sum_{l=0}^{k-1} \frac{1}{l!} \left(\frac{(1-\epsilon)\,c\,t}{m^{1-\alpha}\,i^\alpha} \right)^l \exp\left(-\frac{(1+\epsilon)\,c\,t}{m^{1-\alpha}\,i^\alpha} \right) \\ &\leq m - \sum_{l=0}^{k-1} \frac{1}{l!} \int_{i_\epsilon}^m \left(\frac{(1-\epsilon)\,c\,t}{m^{1-\alpha}\,x^\alpha} \right)^l \exp\left(-\frac{(1+\epsilon)\,c\,t}{m^{1-\alpha}\,x^\alpha} \right) \mathrm{d}x \\ &\quad + \sum_{l=1}^{k-1} \frac{1}{l!} \left(\frac{(1-\epsilon)\,l}{(1+\epsilon)\,e} \right)^l, \end{aligned} \tag{4}$$

where the last term arises from the possibility that the unique maximum of the integrand lies in (i_ϵ, m) for $l = 1, \ldots, k-1$. Similarly, for any $\epsilon \in (0, 1)$ and sufficiently large m,

$$\begin{aligned}\mathrm{E}\overline{S}_k^{(m)}(t) &\geq m - i_\epsilon - \sum_{i=i_\epsilon}^m \sum_{l=0}^{k-1} \frac{1}{l!} \left(\frac{(1+\epsilon)\,c\,t}{m^{1-\alpha}\,i^\alpha} \right)^l \exp\left(-\frac{(1-\epsilon)\,c\,t}{m^{1-\alpha}\,i^\alpha} \right) \\ &\geq m - i_\epsilon - \sum_{l=0}^{k-1} \frac{1}{l!} \int_{i_\epsilon}^{m+1} \left(\frac{(1+\epsilon)\,c\,t}{m^{1-\alpha}\,x^\alpha} \right)^l \exp\left(-\frac{(1-\epsilon)\,c\,t}{m^{1-\alpha}\,x^\alpha} \right) \mathrm{d}x \\ &\quad - \sum_{l=1}^{k-1} \frac{1}{l!} \left(\frac{(1+\epsilon)\,l}{(1-\epsilon)\,e} \right)^l. \end{aligned} \tag{5}$$

Now, let $t = \delta\,m/c$ with constant $\delta > 0$ and change the variable to $u = (1+\epsilon)\,\delta\,m^\alpha/x^\alpha$ in (4) and $u = (1-\epsilon)\,\delta\,m^\alpha/x^\alpha$ in (5) respectively. We then have

$$\mathrm{E}\overline{S}_k^{(m)}\left(\frac{\delta\,m}{c}\right) \leq m - \sum_{l=0}^{k-1} \frac{(1-\epsilon)^l\,\delta^{1/\alpha}\,m}{(1+\epsilon)^{l-1/\alpha}\,\alpha\,l!} \int_{(1+\epsilon)\delta}^{(1+\epsilon)\delta m^\alpha/{i_\epsilon}^\alpha} u^{l-1/\alpha-1}\,e^{-u}\,\mathrm{d}u + o(m), \tag{6}$$

$$\mathrm{E}\overline{S}_k^{(m)}\left(\frac{\delta\,m}{c}\right) \geq m - \sum_{l=0}^{k-1} \frac{(1+\epsilon)^l\,\delta^{1/\alpha}\,m}{(1-\epsilon)^{l-1/\alpha}\,\alpha\,l!} \int_{(1-\epsilon)\delta m^\alpha/(m+1)^\alpha}^{(1-\epsilon)\delta m^\alpha/{i_\epsilon}^\alpha} u^{l-1/\alpha-1}\,e^{-u}\,\mathrm{d}u + o(m). \tag{7}$$

Thus, dividing both sides of (6) and (7) by m, taking $m \to \infty$ and finally $\epsilon \downarrow 0$, we obtain (3). □

Since $S_0^{(m)} = m - \overline{S}_1^{(m)}(t)$ and $S_k^{(m)}(t) = \overline{S}_k^{(m)}(t) - \overline{S}_{k+1}^{(m)}(t)$ a.s. for $k = 1, 2, \ldots$, we can easily derive the limits of $\mathrm{E}S_k^{(m)}(\delta m/c)/m$ as $m \to \infty$ for $k = 0, 1, 2, \ldots$ and $\delta > 0$ from (3) in Lemma 1; that is, Lemma 1 yields the limit of the expected size index distribution for $\alpha < 1$. Contrary to the case of $\alpha > 1$, we can see that this case meets the first definition of LNRE in the sense that $\lim_{m\to\infty} \mathrm{E}S_1^{(m)}(\delta\, m/c) \big/ (\delta\, m/c) > 0$ (so that it does the second definition, too). We now prove the almost sure convergence of the empirical size index distribution.

Theorem 1 *Under* Assumption 2, *we have for any constant* $\delta > 0$ *and* $k = 0, 1, 2, \ldots$,

$$\lim_{m\to\infty} \frac{1}{m}\, S_k^{(m)}\left(\frac{\delta\, m}{c}\right) = \frac{\delta^{1/\alpha}}{\alpha\, k!}\, \Gamma\left(k - \frac{1}{\alpha},\, \delta\right) \quad a.s. \tag{8}$$

Proof Since $\overline{S}_k^{(m)}(t) = \sum_{i=1}^m 1_{\{N_i^{(m)}(t)\geq k\}}$ a.s. and $N_i^{(m)}(t), i = 1, \ldots, m$, are mutually independent, the Chernoff-Hoeffding bound for the sum of 0–1 independent random variables (see, e.g., Mitzenmacher and Upfal (2005, Chapter 4)) implies that, for any $\epsilon > 0$, there exists a $\theta_\epsilon > 0$ such that

$$\mathrm{P}\left(\left|\overline{S}_k^{(m)}(t) - \mathrm{E}\overline{S}_k^{(m)}(t)\right| > \epsilon\, \mathrm{E}\overline{S}_k^{(m)}(t)\right) \leq 2\, e^{-\theta_\epsilon \mathrm{E}\overline{S}_k^{(m)}(t)}. \tag{9}$$

Furthermore, Lemma 1 states that $\mathrm{E}\overline{S}_k^{(m)}(\delta\, m/c) = \Theta(m)$ as $m \to \infty$ under Assumption 2, so that (9) leads to

$$\sum_{m=1}^{\infty} \mathrm{P}\left(\left|\frac{\overline{S}_k^{(m)}(\delta\, m/c)}{\mathrm{E}\overline{S}_k^{(m)}(\delta\, m/c)} - 1\right| > \epsilon\right) < \infty. \tag{10}$$

Hence, the Borel–Cantelli lemma implies that $\overline{S}_k^{(m)}(\delta\, m/c) \big/ \mathrm{E}\overline{S}_k^{(m)}(\delta\, m/c) \to 1$ a.s. as $m \to \infty$, and applying Lemma 1 again, we have for $k = 1, 2, \ldots$,

$$\lim_{m\to\infty} \frac{1}{m}\, \overline{S}_k^{(m)}\left(\frac{\delta\, m}{c}\right) = 1 - \sum_{l=0}^{k-1} \frac{\delta^{1/\alpha}}{\alpha\, l!}\, \Gamma\left(l - \frac{1}{\alpha},\, \delta\right) \quad \text{a.s.},$$

which readily leads to (8). □

Remark 1 Let $\Phi_k(\alpha, \delta)$ denote the right-hand side of (8). Then, as pointed out by an anonymous reviewer, $\Phi_k(\alpha, \delta)$, $\alpha > 0$, $\delta > 0$, is just the mixed Poisson distribution,

$$\Phi_k(\alpha, \delta) = \int_\delta^\infty \frac{u^k}{k!}\, e^{-u}\, \mathrm{d}F(u), \quad k = 0, 1, 2, \ldots,$$

structured by Pareto distribution $\overline{F}(x) = (\delta/x)^{1/\alpha}$, $x \geq \delta$. Thus, $\Phi_k(\alpha, \delta)$, $k = 0, 1, 2, \ldots$, definitely gives a proper distribution on $\mathbb{Z}_+$ and has the asymptotic property $\Phi_k(\alpha, \delta) \sim (\delta^{1/\alpha}/\alpha)\, k^{-1-1/\alpha}$ as $k \to \infty$ even for $\alpha > 0$ (extending the range of α) (see Grandell (1997, Section 8.3)). Furthermore, we can see that, for $\alpha > 0$, $\Phi_k(\alpha, \delta)$ degenerates as

$$\lim_{\delta \downarrow 0} \Phi_k(\alpha, \delta) = \begin{cases} 1, & k = 0, \\ 0, & k = 1, 2, \ldots, \end{cases} \tag{11}$$

which is verified in the Appendix.

To compare our result with that for the case of $\alpha > 1$, we here present the limiting size index distribution in the form of $\lim_{m\to\infty} S_k^{(m)}(\delta\, m/c) \big/ \overline{S}_1^{(m)}(\delta\, m/c)$, $k = 1, 2, \ldots$.

Corollary 1 *Under* Assumption 2*, we have for any constant* $\delta > 0$ *and* $k = 1, 2, \ldots$*,*

$$\lim_{m\to\infty} \frac{S_k^{(m)}(\delta\, m/c)}{\overline{S}_1^{(m)}(\delta\, m/c)} = \frac{\Gamma\,(k - 1/\alpha,\ \delta)}{\alpha\, k!\ \left[\Gamma\,(1 - 1/\alpha,\ \delta) + \delta^{-1/\alpha}\,(1 - e^{-\delta})\right]} \quad a.s. \tag{12}$$

Remark 2 Let $\Psi_k(\alpha, \delta)$ denote the right-hand side of (12). We can then see that $\Psi_k(\alpha, \delta)$, $k = 1, 2, \ldots$, is the zero-truncated distribution of $\Phi_k(\alpha, \delta)$, which is given on the right-hand side of (8). Thus, from Remark 1, it also gives a proper distribution on $\mathbb{N}$ with $\Psi_k(\alpha, \delta) = \Theta(k^{-1-1/\alpha})$ as $k \to \infty$ even for $\alpha > 0$ (extending the range of α). Furthermore, we can show that, when $\alpha > 1$, $\lim_{\delta\downarrow 0} \Psi_k(\alpha, \delta)$, $k = 1, 2, \ldots$, reduces to the KRS distribution $\Psi_k(\alpha)$, $k = 1, 2, \ldots$, which is given on the right-hand side of (2), but when $\alpha \in (0, 1]$, it degenerates as

$$\lim_{\delta \downarrow 0} \Psi_k(\alpha, \delta) = \begin{cases} 1, & k = 1, \\ 0, & k = 2, 3, \ldots, \end{cases} \tag{13}$$

which is verified in the Appendix. It should be noted that Evert (2004) derives similar formulas to (8) and (12) by using a different approach.

2.4 Case of $\alpha = 1$

In the final subsection of the independent throw model, we consider the case of $\alpha = 1$ and $m < \infty$. The frequency distribution is given by the following.

Assumption 3 For any $\epsilon > 0$, there exists an integer $i_\epsilon > 0$ such that, for all m and i satisfying $m \geq i \geq i_\epsilon$, inequality $(1 - \epsilon)\, c/(i\ \ln m) \leq p_i^{(m)} \leq (1 + \epsilon)\, c/(i\ \ln m)$ holds with $c > 0$.

Similar to Assumption 2 in the preceding subsection, Assumption 3 represents the Zipf-type distribution with $\alpha = 1$. Indeed, if $p_i^{(m)} = c_m/i$ for $i = 1, \ldots, m$ with the normalization constant c_m, we have $c_m = (\sum_{i=1}^m 1/i)^{-1} \sim 1/\ln m$ as $m \to \infty$ and Assumption 3 is fulfilled with $c = 1$. As in the preceding subsection, we first provide the asymptotics of the expected size indices.

Lemma 2 *Under* Assumption 3, *for any fixed constant* $\delta > 0$ *and* $k = 1, 2, \ldots$,

$$\mathrm{E}\overline{S}_k^{(m)}\left(\frac{\delta\, m\, \ln m}{c}\right) \sim m\left[1 - \sum_{l=0}^{k-1} \frac{\delta}{l!}\, \Gamma(l-1,\, \delta)\right] \quad \text{as } m \to \infty. \tag{14}$$

Proof The proof is similar to that of Lemma 1. The differences are in that, under Assumption 3, inequalities (4) and (5) are respectively replaced by

$$\begin{aligned} \mathrm{E}\overline{S}_k^{(m)}(t) \le m &- \sum_{l=0}^{k-1} \frac{1}{l!} \int_{i_\epsilon}^{m} \left(\frac{(1-\epsilon)\, c\, t}{x\, \ln m}\right)^l \exp\left(-\frac{(1+\epsilon)\, c\, t}{x\, \ln m}\right) \mathrm{d}x \\ &+ \sum_{l=1}^{k-1} \frac{1}{l!} \left(\frac{(1-\epsilon)\, l}{(1+\epsilon)\, e}\right)^l, \end{aligned} \tag{15}$$

$$\begin{aligned} \mathrm{E}\overline{S}_k^{(m)}(t) \ge m - i_\epsilon &- \sum_{l=0}^{k-1} \frac{1}{l!} \int_{i_\epsilon}^{m+1} \left(\frac{(1+\epsilon)\, c\, t}{x\, \ln m}\right)^l \exp\left(-\frac{(1-\epsilon)\, c\, t}{x\, \ln m}\right) \mathrm{d}x \\ &- \sum_{l=1}^{k-1} \frac{1}{l!} \left(\frac{(1+\epsilon)\, l}{(1-\epsilon)\, e}\right)^l. \end{aligned} \tag{16}$$

Thus, letting $t = (\delta\, m\, \ln m)/c$ with $\delta > 0$ and changing the variable to $u = (1+\epsilon)\, \delta\, m/x$ in (15) and $u = (1-\epsilon)\, \delta\, m/x$ in (16) respectively, we have

$$\mathrm{E}\overline{S}_k^{(m)}\left(\frac{\delta\, m\, \ln m}{c}\right) \le m - \sum_{l=0}^{k-1} \frac{(1-\epsilon)^l\, \delta\, m}{(1+\epsilon)^{l-1}\, l!} \int_{(1+\epsilon)\delta}^{(1+\epsilon)\delta m/i_\epsilon} u^{l-2}\, e^{-u}\, \mathrm{d}u + o(m),$$

$$\mathrm{E}\overline{S}_k^{(m)}\left(\frac{\delta\, m\, \ln m}{c}\right) \ge m - \sum_{l=0}^{k-1} \frac{(1+\epsilon)^l\, \delta\, m}{(1-\epsilon)^{l-1}\, l!} \int_{(1-\epsilon)\delta m/(m+1)}^{(1-\epsilon)\delta m/i_\epsilon} u^{l-2}\, e^{-u}\, \mathrm{d}u + o(m).$$

Hence, dividing both sides by m, taking $m \to \infty$ and then $\epsilon \downarrow 0$, we obtain (14). □

Note here that the right-hand side of (14) in Lemma 2 is just the version of $\alpha = 1$ on the right-hand side of (3) in Lemma 1 while the functions $t = t(m), m \in \mathbb{N}$, are slightly different for the two cases. Concerning the LNRE property, however, we can see that this case has rather a similar feature to the case of $\alpha > 1$; that is, it does not meet the first definition of LNRE in the sense that $\lim_{m\to\infty} \mathrm{E}S_1^{(m)}(\delta\, m\, \ln m/c)/(\delta\, m\, \ln m/c) = 0$ but does the second definition in the sense that $\lim_{m\to\infty} \mathrm{E}\overline{S}_1^{(m)}(\delta\, m\, \ln m/c) = \infty$ and $\lim_{m\to\infty} \mathrm{E}S_1^{(m)}(\delta\, m\, \ln m/c)/\mathrm{E}\overline{S}_1^{(m)}(\delta\, m\, \ln m/c) > 0$. Applying Lemma 2, a similar procedure to the proof of Theorem 1 yields the following.

Theorem 2 *Under* Assumption 3, *for any fixed constant* $\delta > 0$ *and* $k = 0, 1, 2, \ldots$,

$$\lim_{m\to\infty} \frac{1}{m}\, S_k^{(m)}\left(\frac{\delta\, m\, \ln m}{c}\right) = \frac{\delta}{k!}\, \Gamma(k-1,\, \delta) \quad a.s. \tag{17}$$

Proof The difference from the proof of Theorem 1 is just that, applying Lemma 2 under Assumption 3 (instead of Lemma 1 under Assumption 2), (10) is replaced by

$$\sum_{m=1}^{\infty} \mathrm{P}\left(\left|\frac{\overline{S}_k^{(m)}(\delta\, m\, \ln m/c)}{\mathrm{E}\overline{S}_k^{(m)}(\delta\, m\, \ln m/c)} - 1\right| > \epsilon\right) < \infty.$$

The remaining procedure is quite the same. □

As seen in Remark 1, the right-hand side of (17) in Theorem 2 gives a proper distribution on $\mathbb{Z}_+$. In the case of $\alpha = 1$, of course, the following holds.

Corollary 2 *Under* Assumption 3, *we have for any fixed constant* $\delta > 0$ *and* $k = 1, 2, \ldots$,

$$\lim_{m\to\infty} \frac{S_k^{(m)}\,((\delta\, m\, \ln m)/c)}{\overline{S}_1^{(m)}\,((\delta\, m\, \ln m)/c)} = \frac{\Gamma(k-1,\, \delta)}{k!\,\left[\Gamma(0,\, \delta) + \delta^{-1}\,(1 - e^{-\delta})\right]} \quad a.s. \tag{18}$$

3 Dependent throw model

We here extend the results in the preceding section to the dependent throw model. We will find that the derived limiting size index distributions are still valid even in the case where the choices of bins are dependent in some general sense.

3.1 Model description

As in the independent throw model, balls are thrown according to a homogeneous Poisson process with intensity 1, but the choice of a bin at each throw is governed by the following doubly stochastic structure. Let $\{\boldsymbol{q}^{(m)}(t)\}_{t\ge 0}$ denote a stochastic process on $[0, 1]^m$ independent of the Poisson process $\{N(t)\}_{t\ge 0}$ such that $\boldsymbol{q}^{(m)}(t) = (q_1^{(m)}(t), \ldots, q_m^{(m)}(t))$ satisfies $\sum_{i=1}^m q_i^{(m)} = 1$ a.s. for $t \ge 0$. We assume that $\mathrm{E}q_i^{(m)}(t) = p_i^{(m)}$ for $i = 1, \ldots, m$ and $t \ge 0$. Let T_n, $n = 1, 2, \ldots$, denote the nth point of the Poisson process $\{N(t)\}_{t\ge 0}$ at which the nth ball is thrown. Then, once $\boldsymbol{q}^{(m)}(T_n)$ is given at time T_n, $n = 1, 2, \ldots$, the chosen bin B_n is conditionally independent of $\{B_l\}_{l\ne n}$ and $\{\boldsymbol{q}^{(m)}(t)\}_{t\ne T_n}$, and the nth ball enters bin i $(= 1, \ldots, m)$ with conditional probability $q_i^{(m)}(T_n) = \mathrm{P}(B_n = i \,|\, \boldsymbol{q}^{(m)}(T_n))$. Within this setting, each $\{N_i^{(m)}(t)\}_{t\ge 0}$, $i = 1, \ldots, m$, is a Cox (doubly stochastic Poisson) process with random intensity process $\{q_i^{(m)}(t)\}_{t\ge 0}$, and $N_i^{(m)}(t)$, $i = 1, \ldots, m$, are mutually conditionally independent given $\{\boldsymbol{q}^{(m)}(s)\}_{s\le t}$; that is,

$$\mathrm{P}\left(N_1^{(m)}(t) = k_1, \ldots, N_m^{(m)}(t) = k_m \mid \boldsymbol{q}^{(m)}(s), s \le t\right)$$
$$= \prod_{i=1}^{m} \frac{\left[Q_i^{(m)}(t)\right]^{k_i}}{k_i!}\, e^{-Q_i^{(m)}(t)} \quad \text{a.s.,} \tag{19}$$

where $Q_i^{(m)}(t) = \int_0^t q_i^{(m)}(s)\,\mathrm{d}s$ for $i = 1, \ldots, m$.

3.2 Case of $\alpha > 1$

We here extend Propositions 1 and 2 in Sect. 2.2 to the dependent throw model. As in Section 2.2, we suppress the superscript "(∞)" and write, for example, $\boldsymbol{q}(t)$ for $\boldsymbol{q}^{(\infty)}(t)$ and so on. Assuming that $\{\boldsymbol{q}(t)\}_{t\geq 0}$ is ergodic, the ergodic theorem implies that $Q_i(t)/t \to p_i$ a.s. as $t \to \infty$ for $i = 1, 2, \ldots$. Now, for any $t > 0$ and $\epsilon > 0$, we define $A_\epsilon(t) \in \mathcal{F}$ such that

$$A_\epsilon(t) = \left\{ \omega \in \Omega \;\middle|\; \sup_{i\in\mathbb{N}} \left| \frac{Q_i(\omega, t)}{p_i\, t} - 1 \right| \leq \epsilon \right\}. \tag{20}$$

Note that the ergodicity implies $\mathrm{P}(A_\epsilon(t)) \to 1$ as $t \to \infty$. We first show the asymptotics of the expected size indices, where some convergence speed of $\mathrm{P}\left(A_\epsilon(t)^c\right) \to 0$ as $t \to \infty$ is required.

Lemma 3 *Under* Assumption 1, *if* $\mathrm{P}(A_\epsilon(t)^c) = o(t^{-1+1/\alpha})$, *then* (1) *holds for* $k = 1, 2, \ldots$.

Proof We first consider the case of $k = 1$; that is, we verify that, under the condition of the lemma,

$$\mathrm{E}\overline{S}_1(t) \sim c^{1/\alpha}\, t^{1/\alpha}\, \Gamma\left(1 - \frac{1}{\alpha}\right) \quad \text{as } t \to \infty. \tag{21}$$

Since $\mathrm{E}(\overline{S}_1(t) \mid \boldsymbol{q}(s), s \leq t) = \sum_{i=1}^{\infty} \mathrm{P}(N_i(t) \geq 1 \mid q_i(s), s \leq t)$ a.s., we have from (19) that

$$\mathrm{E}(\overline{S}_1(t) \mid \boldsymbol{q}(s), s \leq t) = \sum_{i=1}^{\infty} \left[1 - e^{Q_i(t)}\right] \quad \text{a.s.} \tag{22}$$

Here, for any $\epsilon \in (0, 1)$ and $i = 1, 2, \ldots$, we have $(1-\epsilon)\, p_i\, t \leq Q_i(t) \leq (1+\epsilon)\, p_i\, t$ on $A_\epsilon(t)$ in (20), and thus, under Assumption 1,

$$\begin{aligned} \mathrm{E}(\overline{S}_1(t) \mid \boldsymbol{q}(s), s \leq t) &\leq \sum_{i=1}^{\infty} \left[1 - e^{-(1+\epsilon) p_i t}\right] + t\, 1_{A_\epsilon(t)^c} \\ &\leq i_\epsilon + \sum_{i=i_\epsilon+1}^{\infty} \left[1 - \exp\left(-\frac{(1+\epsilon)^2\, c\, t}{i^\alpha}\right)\right] + t\, 1_{A_\epsilon(t)^c} \\ &\leq i_\epsilon + \int_{i_\epsilon}^{\infty} \left[1 - \exp\left(-\frac{(1+\epsilon)^2\, c\, t}{x^\alpha}\right)\right] \mathrm{d}x + t\, 1_{A_\epsilon(t)^c} \quad \text{a.s.,} \end{aligned} \tag{23}$$

where the first inequality follows from $\overline{S}_1(t) \leq N(t)$ a.s. and $\mathrm{E}(N(t) \mid \boldsymbol{q}(s), s \leq t) = \mathrm{E}N(t) = t$, and the last inequality follows from the fact that $1 - \exp\left(-(1+\epsilon)^2\, c\, t/i^\alpha\right)$

is decreasing in $i > 1$. For the integral in the last expression of (23), changing the variable to $u = (1+\epsilon)^2\, c\, t/x^\alpha$ and then integrating it by parts, we have

$$\int_{i_\epsilon}^{\infty}\left[1-\exp\left(-\frac{(1+\epsilon)^2\, c\, t}{x^\alpha}\right)\right]\mathrm{d}x = -i_\epsilon\left[1-\exp\left(-\frac{(1+\epsilon)^2\, c\, t}{{i_\epsilon}^\alpha}\right)\right]$$
$$+(1+\epsilon)^{2/\alpha}\, c^{1/\alpha}\, t^{1/\alpha}\int_0^{(1+\epsilon)^2 ct/{i_\epsilon}^\alpha} u^{-1/\alpha}\, e^{-u}\,\mathrm{d}u.$$

Therefore, after substituting this into (23), taking the expectation on both its sides, dividing it by $t^{1/\alpha}$, then taking $t \to \infty$ and $\epsilon \downarrow 0$, we have under the condition that $\mathrm{P}(A_\epsilon(t)^c) = o(t^{-1+1/\alpha})$,

$$\limsup_{t\to\infty}\frac{\mathrm{E}\overline{S}_1(t)}{t^{1/\alpha}} \le c^{1/\alpha}\,\Gamma\left(1-\frac{1}{\alpha}\right).$$

To verify (21), we still need to show the asymptotic lower bound. Similar to the above argument, we have from (22) that, under Assumption 1,

$$\begin{aligned}\mathrm{E}(\overline{S}_1(t)\mid \boldsymbol{q}(s), s\le t) &\ge \sum_{i=1}^{\infty}[1-e^{-(1-\epsilon)p_i t}]\, 1_{A_\epsilon(t)}\\ &\ge \sum_{i=i_\epsilon}^{\infty}\left[1-\exp\left(-\frac{(1-\epsilon)^2\, c\, t}{i^\alpha}\right)\right] 1_{A_\epsilon(t)}\\ &\ge \int_{i_\epsilon}^{\infty}\left[1-\exp\left(-\frac{(1-\epsilon)^2\, c\, t}{x^\alpha}\right)\right]\mathrm{d}x\; 1_{A_\epsilon(t)} \quad \text{a.s.}\end{aligned}$$

Hence, similar to obtaining the asymptotic upper bound, we have

$$\liminf_{t\to\infty}\frac{\mathrm{E}\overline{S}_1(t)}{t^{1/\alpha}} \ge c^{1/\alpha}\,\Gamma\left(1-\frac{1}{\alpha}\right),$$

since $\mathrm{P}(A_\epsilon(t)) \to 1$ as $t\to\infty$, so that (21) is obtained.

We next show that, under the condition of the lemma,

$$\mathrm{E}S_k(t) \sim \frac{c^{1/\alpha}}{\alpha\, k!}\, t^{1/\alpha}\,\Gamma\left(k-\frac{1}{\alpha}\right) \quad \text{as } t\to\infty,\quad k=1,2,\ldots. \tag{24}$$

By using a similar argument to the above, we have from (19) and (20) that, under Assumption 1, for any $\epsilon \in (0, 1)$,

$$\begin{aligned} \mathrm{E}\,(S_k(t) \mid \boldsymbol{q}(s), s \le t) &\le \sum_{i=1}^{\infty} \frac{[(1+\epsilon)\, p_i\, t]^k}{k!}\, e^{-(1-\epsilon) p_i t} + t\, 1_{A_\epsilon(t)^c} \\ &\le i_\epsilon + \frac{1}{k!} \sum_{i=i_\epsilon+1}^{\infty} \left[\frac{(1+\epsilon)^2\, c\, t}{i^\alpha}\right]^k \exp\left(-\frac{(1-\epsilon)^2\, c\, t}{i^\alpha}\right) \\ &\quad + t\, 1_{A_\epsilon(t)^c} \quad \text{a.s.} \end{aligned} \tag{25}$$

Here, the summand of (25) is decreasing in sufficiently large i and we can choose $i_\epsilon > 0$ such that it is decreasing in $i \ge i_\epsilon$. Then,

$$\begin{aligned} \mathrm{E}\,(S_k(t) \mid \boldsymbol{q}(s), s \le t) &\le i_\epsilon + \frac{1}{k!} \int_{i_\epsilon+1}^{\infty} \left[\frac{(1+\epsilon)^2\, c\, t}{x^\alpha}\right]^k \exp\left(-\frac{(1-\epsilon)^2\, c\, t}{x^\alpha}\right) \mathrm{d}x \\ &\quad + t\, 1_{A_\epsilon(t)^c} \\ &= i_\epsilon + \frac{(1+\epsilon)^{2k}\, c^{1/\alpha}\, t^{1/\alpha}}{(1-\epsilon)^{2(k-1/\alpha)}\, \alpha\, k!} \int_0^{(1-\epsilon)^2 ct/i_\epsilon{}^\alpha} u^{k-1/\alpha-1}\, e^{-u}\, \mathrm{d}u \\ &\quad + t\, 1_{A_\epsilon(t)^c} \quad \text{a.s.,} \end{aligned} \tag{26}$$

where the last equality follows from changing the variable to $u = (1-\epsilon)^2\, c\, t/x^\alpha$. Thus, taking the expectation on both sides of (26), dividing it by $t^{1/\alpha}$, further taking $t \to \infty$ and $\epsilon \downarrow 0$, we have under the condition that $\mathrm{P}(A_\epsilon(t)^c) = o(t^{-1+1/\alpha})$,

$$\limsup_{t\to\infty} \frac{\mathrm{E} S_k(t)}{t^{1/\alpha}} \le \frac{c^{1/\alpha}}{\alpha\, k!}\, \Gamma\left(k - \frac{1}{\alpha}\right).$$

Similarly, we have under Assumption 1,

$$\begin{aligned} \mathrm{E}\,(S_k(t) \mid \boldsymbol{q}(s), s \le t) &\ge \sum_{i=1}^{\infty} \frac{[(1-\epsilon)\, p_i\, t]^k}{k!}\, e^{-(1+\epsilon) p_i t}\, 1_{A_\epsilon(t)} \\ &\ge \frac{1}{k!} \sum_{i=i_\epsilon}^{\infty} \left[\frac{(1-\epsilon)^2\, c\, t}{i^\alpha}\right]^k \exp\left(-\frac{(1+\epsilon)^2\, c\, t}{i^\alpha}\right) 1_{A_\epsilon(t)} \quad \text{a.s.,} \end{aligned}$$

from which, a similar procedure to the above yields

$$\liminf_{t\to\infty} \frac{\mathrm{E} S_k(t)}{t^{1/\alpha}} \ge \frac{c^{1/\alpha}}{\alpha\, k!}\, \Gamma\left(k - \frac{1}{\alpha}\right),$$

since $\mathrm{P}(A_\epsilon(t)) \to 1$ as $t \to \infty$, and hence, (24) is obtained. Finally, (1) is immediate from (21) and (24). □

By using Lemma 3 and the same argument as that after Proposition 1, we can derive the KRS distribution $\Psi_k(\alpha)$, $k = 1, 2, \ldots$, given on the right-hand side of (2), as the limit of $\mathrm{E}S_k(t)/\mathrm{E}\overline{S}_1(t)$ as $t \to \infty$ under the same condition as that of the lemma. Now, let us demonstrate the almost sure convergence under a somewhat stronger condition.

Theorem 3 *Under* Assumption 1, *if* $\sum_{n=1}^{\infty} \mathrm{P}(A_\epsilon(n)^c) < \infty$, *then* (2) *holds for* $k = 1, 2, \ldots$.

To prove Theorem 3, we use the next lemma, where after this, $\{N_i^\dagger(t)\}_{t\ge 0}$, $i = 1, 2, \ldots$, denotes a homogeneous Poisson process with intensity p_i and $\overline{S}_k^\dagger(t) = \sum_{i=1}^{\infty} 1_{\{N_i^\dagger(t)\ge k\}}$ for $k = 1, 2, \ldots$. That is, $\{N_i^\dagger(t)\}_{t\ge 0}$ and $\overline{S}_k^\dagger(t)$ considered here are respectively nothing but $\{N_i(t)\}_{t\ge 0}$ and $\overline{S}_k(t)$ for the independent throw model considered in the preceding section.

Lemma 4 *For any* $\epsilon > 0$, *there exists a* $\theta_\epsilon > 0$ *such that*

$$\mathrm{P}\left(\overline{S}_k(t) > (1+\epsilon)\,\mathrm{E}\overline{S}_k^\dagger((1+\epsilon)\,t)\right) \le e^{-\theta_\epsilon \mathrm{E}\overline{S}_k^\dagger((1+\epsilon)t)} + \mathrm{P}(A_\epsilon(t)^c), \tag{27}$$

$$\mathrm{P}\left(\overline{S}_k(t) < (1-\epsilon)\,\mathrm{E}\overline{S}_k^\dagger((1-\epsilon)\,t)\right) \le e^{-\theta_\epsilon \mathrm{E}\overline{S}_k^\dagger((1-\epsilon)\,t)} + \mathrm{P}(A_\epsilon(t)^c). \tag{28}$$

Lemma 4 plays the role of (9) in the case of dependent throws. Equations (27) and (28) are also available when $m < \infty$ and are indeed exploited in the following subsections.

Proof It is clear that

$$\begin{aligned}
&\mathrm{P}\left(\overline{S}_k(t) > (1+\epsilon)\,\mathrm{E}\overline{S}_k^\dagger((1+\epsilon)\,t)\right)\\
&\quad\le \mathrm{P}\left(\left\{\overline{S}_k(t) > (1+\epsilon)\,\mathrm{E}\overline{S}_k^\dagger((1+\epsilon)\,t)\right\} \cap A_\epsilon(t)\right) + \mathrm{P}(A_\epsilon(t)^c),
\end{aligned}$$

and we consider the first term on the right-hand side above. Since a Poisson random variable is stochastically increasing in its mean, Theorem 1.A.14 in Shaked and Shanthikumar (2007) implies that

$$N_i(t)\,1_{\{Q_i(t)\le(1+\epsilon)p_i t\}} \le_{\mathrm{st}} N_i^\dagger((1+\epsilon)\,t)\,1_{\{Q_i(t)\le(1+\epsilon)p_i t\}}, \quad i = 1, 2, \ldots,$$

where "$\le_{\mathrm{st}}$" represents the usual stochastic order (see, e.g., Müller and Stoyan (2002) or Shaked and Shanthikumar (2007)). Thus, since $\overline{S}_k(t)$ is a.s. nondecreasing in $N_i(t)$, $i = 1, 2, \ldots$, we have $\overline{S}_k(t)\,1_{A_\epsilon(t)} \le_{\mathrm{st}} \overline{S}_k^\dagger((1+\epsilon)\,t)\,1_{A_\epsilon(t)}$, which implies that

$$\begin{aligned}
&\mathrm{P}\left(\left\{\overline{S}_k(t) > (1+\epsilon)\,\mathrm{E}\overline{S}_k^\dagger((1+\epsilon)\,t)\right\} \cap A_\epsilon(t)\right)\\
&\quad\le \mathrm{P}\left(\overline{S}_k^\dagger((1+\epsilon)\,t) > (1+\epsilon)\,\mathrm{E}\overline{S}_k^\dagger((1+\epsilon)\,t)\right).
\end{aligned}$$

Hence, the Chernoff–Hoeffding bound for the sum of 0–1 independent random variables yields (27). Inequality (28) is similarly verified. □

Proof (Proof of Theorem 3) By using (27) in Lemma 4, we have

$$P\left(\frac{\overline{S}_k(t)}{E\overline{S}_k^\dagger((1+\epsilon)\,t)} - 1 > \epsilon\right) \le e^{-\theta_\epsilon E\overline{S}_k^\dagger((1+\epsilon)t)} + P\left(A_\epsilon(t)^c\right). \tag{29}$$

Here, since $\overline{S}_k^\dagger((1+\epsilon)\,t)$ is nothing but $\overline{S}_k((1+\epsilon)\,t)$ for the independent throw model considered in the preceding section, Proposition 1 implies that $E\overline{S}_k^\dagger((1+\epsilon)\,t) = \Theta(t^{1/\alpha})$ as $t \to \infty$, so that, under the condition of the theorem, (29) leads to

$$\sum_{n=1}^{\infty} P\left(\frac{\overline{S}_k(n)}{E\overline{S}_k^\dagger((1+\epsilon)\,n)} - 1 > \epsilon\right) < \infty.$$

Therefore, the Borel–Cantelli lemma implies that

$$\limsup_{n\to\infty} \frac{\overline{S}_k(n)}{E\overline{S}_k^\dagger((1+\epsilon)\,n)} \le 1+\epsilon \quad \text{a.s.}$$

A similar argument based on (28) yields

$$\liminf_{n\to\infty} \frac{\overline{S}_k(n)}{E\overline{S}_k^\dagger((1-\epsilon)\,n)} \ge 1-\epsilon \quad \text{a.s.}$$

Hence, applying Proposition 1 again and taking $\epsilon \downarrow 0$, we have $\overline{S}_k(n) \sim E\overline{S}_k^\dagger(n)$ a.s. as $n \to \infty$ on $n \in \mathbb{N}$, which leads to (2) in the case where t goes to infinity on $t \in \mathbb{N}$, but the extension to that on $t \in \mathbb{R}$ is easy. □

3.3 Case of $\alpha < 1$

We extend the results in Sect. 2.3. To this end, we define for $t \ge 0$ and $\epsilon > 0$,

$$A_\epsilon^{(m)}(t) = \left\{\omega \in \Omega \;\middle|\; \max_{i\in\{1,\ldots,m\}} \left|\frac{Q_i^{(m)}(\omega,t)}{p_i^{(m)}\,t} - 1\right| \le \epsilon\right\}. \tag{30}$$

As in the independent throw model, we first show the asymptotics of the expected size indices.

Lemma 5 *Under* Assumption 2, *if* $P(A_\epsilon^{(m)}(\delta\,m/c)) \to 1$ *as* $m \to \infty$ *for some* $\delta > 0$, *then* (3) *holds for such* δ *and* $k = 1, 2, \ldots$.

Recall that, in the case of $\alpha > 1$, we assume that $\{q^{(\infty)}(t)\}_{t\geq 0}$ is ergodic, which implies that $\mathrm{P}(A_\epsilon^{(\infty)}(t)) \to 1$ as $t \to \infty$. Furthermore, some speed in this convergence is required in Lemma 3 and Theorem 3. Here, unlike the case of $\alpha > 1$, $\{q^{(m)}(\delta\, m/c)\}_{m\in\mathbb{N}}$ is no longer ergodic and $\mathrm{P}(A_\epsilon^{(m)}(\delta\, m/c)^c)$ is just required to vanish as $m \to \infty$ to derive the asymptotics of the expected size indices, where the vanishing speed is not required.

Proof By using (19), we have

$$\mathrm{E}\left(\overline{S}_k^{(m)} \mid \boldsymbol{q}^{(m)}(s), s \leq t\right) = \sum_{i=1}^{m} \left[1 - \sum_{l=0}^{k-1} \frac{Q_i^{(m)}(t)^l}{l!}\, e^{-Q_i^{(m)}(t)}\right] \quad \text{a.s.}$$

For any $\epsilon > 0$, we have $(1-\epsilon)\, p_i^{(m)}\, t \leq Q_i^{(m)}(t) \leq (1+\epsilon)\, p_i^{(m)}\, t$ on $A_\epsilon^{(m)}(t)$ in (30), and thus, under Assumption 2,

$$\begin{aligned}&\mathrm{E}\left(\overline{S}_k^{(m)} \mid \boldsymbol{q}^{(m)}(s), s \leq t\right)\\ &\quad\leq m - \sum_{i=i_\epsilon}^{m}\sum_{l=0}^{k-1} \frac{1}{l!}\left(\frac{(1-\epsilon)^2\, c\, t}{m^{1-\alpha}\, i^\alpha}\right)^l \exp\left(-\frac{(1+\epsilon)^2\, c\, t}{m^{1-\alpha}\, i^\alpha}\right) 1_{A_\epsilon^{(m)}(t)} \quad \text{a.s.,} \qquad (31)\end{aligned}$$

$$\begin{aligned}&\mathrm{E}\left(\overline{S}_k^{(m)} \mid \boldsymbol{q}^{(m)}(s), s \leq t\right)\\ &\quad\geq (m - i_\epsilon)\, 1_{A_\epsilon^{(m)}(t)} - \sum_{i=i_\epsilon}^{m}\sum_{l=0}^{k-1} \frac{1}{l!}\left(\frac{(1+\epsilon)^2\, c\, t}{m^{1-\alpha}\, i^\alpha}\right)^l \exp\left(-\frac{(1-\epsilon)^2\, c\, t}{m^{1-\alpha}\, i^\alpha}\right) \quad \text{a.s.}\end{aligned} \qquad (32)$$

Take the expectation on both sides of (31) and (32) and let $t = \delta\, m/c$ with $\delta > 0$. Then, since $\mathrm{P}(A_\epsilon^{(m)}(\delta\, m/c)) \to 1$ as $m \to \infty$ under the condition of the lemma, the remaining procedure is almost the same as that in the proof of Lemma 1. □

Next, we extend Theorem 1 and Corollary 1, where some vanishing speed of $\mathrm{P}(A_\epsilon^{(m)}(\delta\, m/c)^c) \to 0$ as $m \to \infty$ is required.

Theorem 4 *Under* Assumption 2, *if* $\sum_{m=1}^\infty \mathrm{P}(A_\epsilon^{(m)}(\delta\, m/c)^c) < \infty$ *for some* $\delta > 0$, *then* (8) *holds for such* δ *and* $k = 0, 1, 2, \ldots$, *and also* (12) *holds for* $k = 1, 2, \ldots$.

Proof As in the proof of Theorem 3, let $\overline{S}_k^{\dagger(m)}(t)$, $k = 1, 2, \ldots$, just represent $\overline{S}_k^{(m)}(t)$ for the independent throw model. Applying (27) in Lemma 4, we have

$$\mathrm{P}\left(\frac{\overline{S}_k^{(m)}(\delta\, m/c)}{\mathrm{E}\overline{S}_k^{\dagger(m)}((1+\epsilon)\,\delta\, m/c)} - 1 > \epsilon\right) \leq e^{-\theta_\epsilon \mathrm{E}\overline{S}_k^{\dagger(m)}((1+\epsilon)\delta m/c)} + \mathrm{P}\left(A_\epsilon(\delta\, m/c)^c\right).$$

Here, (3) in Lemma 1 implies that $\mathrm{E}\overline{S}_k^{\dagger(m)}((1+\epsilon)\,\delta\,m/c) = \Theta(m)$ as $m \to \infty$, so that, we have under the condition of the theorem,

$$\sum_{m=1}^{\infty} \mathrm{P}\left(\frac{\overline{S}_k^{(m)}(\delta\,m/c)}{\mathrm{E}\overline{S}_k^{\dagger(m)}((1+\epsilon)\,\delta\,m/c)} - 1 > \epsilon \right) < \infty.$$

Therefore, the Borel–Cantelli lemma yields that

$$\limsup_{m\to\infty} \frac{\overline{S}_k^{(m)}(\delta\,m/c)}{\mathrm{E}\overline{S}_k^{\dagger(m)}((1+\epsilon)\,\delta\,m/c)} \le 1+\epsilon \quad \text{a.s.}$$

Similarly, by applying (28) in Lemma 4, we have

$$\liminf_{m\to\infty} \frac{\overline{S}_k^{(m)}(\delta\,m/c)}{\mathrm{E}\overline{S}_k^{\dagger(m)}((1-\epsilon)\,\delta\,m/c)} \ge 1-\epsilon \quad \text{a.s.}$$

Hence, applying Lemma 1 and taking $\epsilon \downarrow 0$, we have $\overline{S}_k^{(m)}(\delta\,m/c) \sim \mathrm{E}\overline{S}_k^{\dagger(m)}(\delta\,m/c)$ a.s. as $m \to \infty$, which implies (8) and (12). □

3.4 Case of $\alpha = 1$

Even in the case of $\alpha = 1$, a similar argument to that in the preceding subsections leads to the following results. As the differences in the proofs are just that we apply Assumption 3 instead of Assumption 2 and replace $t = \delta\,m/c$ with $t = (\delta\,m\,\ln m)/c$, the proofs have been omitted.

Lemma 6 *Under* Assumption 3, *if* $\mathrm{P}(A_\epsilon^{(m)}((\delta\,m\,\ln m)/c)) \to 1$ *as* $m \to \infty$ *for some* $\delta > 0$, *then* (14) *holds for such* δ *and* $k = 1, 2, \ldots$.

Theorem 5 *Under* Assumption 3, *if* $\sum_{m=1}^{\infty} \mathrm{P}(A_\epsilon^{(m)}((\delta\,m\,\ln m)/c)^c) < \infty$ *for some* $\delta > 0$, *then* (17) *holds for such* δ *and* $k = 0, 1, 2, \ldots$, *and also* (18) *holds for* $k = 1, 2, \ldots$.

4 Simulation experiments

We here discuss our validation of the theoretical results derived in the previous sections done through simulation experiments; that is, we compare the limiting size index distributions derived in Sects. 2 and 3 with the estimates of the empirical size index distributions obtained with the simulations. Software R for statistical computing (R Development Core Team 2007) was used for the simulation experiments. In each experiment, 100 independent replicates of sample paths with length $t = 1{,}000$ were executed, and the means and the 95% confidence intervals were calculated. In each figure below, both axes are logarithmically scaled. From the experimental results, we will find not only that our analysis is valid but also that the derived limiting distributions well approximate the empirical size index distributions even in a relatively short period with about 1,000 balls.

4.1 Independent throw model

Example 1 (Case of $\alpha > 1$) In the first example, we examine the independent throw model with $\alpha > 1$. In the experiment, the frequency distribution was given by $p_i = c/i^\alpha$, $i = 1, 2, \ldots$, and two cases of $\alpha = 1.1$ and $\alpha = 1.5$ were executed. The value of the normalization constant c was calculated as $c^{-1} = \sum_{i=1}^{i_0} 1/i^\alpha + {i_0}^{-\alpha+1}/(\alpha - 1)$ with $i_0 = 2 \times 10^7$, which is based on $\sum_{i>i_0} 1/i^\alpha \sim {i_0}^{-\alpha+1}/(\alpha - 1)$ as $i_0 \to \infty$, and the empirical size index distribution $S_k(t)/\overline{S}_1(t)$, $k = 1, 2, \ldots$, with $t = 1{,}000$ was estimated by simulation. To compare the simulation estimates with the theoretical result and distinguish the two kinds of plots, we extend the range of the function $\Psi_k(\alpha)$, $k = 1, 2, \ldots$, on the right-hand side of (2) in Proposition 2 to $\mathbb{R}_+$; that is, applying $k! = \Gamma(k + 1)$, the curves of the function,

$$\Psi(x; \alpha) = \frac{\Gamma(x - 1/\alpha)}{\alpha\, \Gamma(x + 1)\, \Gamma(1 - 1/\alpha)}, \quad x > 0, \tag{33}$$

are evaluated. The value of $\Psi(x; \alpha)$ above, of course, coincides with that of $\Psi_x(\alpha)$ at $x = 1, 2, \ldots$. The results are given in Fig. 1, where we can see good agreement in the theoretical results with the simulation estimates for both cases of $\alpha = 1.1$ and $\alpha = 1.5$.

Example 2 (Case of $\alpha < 1$) In the second example, let us examine the independent throw model with $\alpha < 1$. In the experiment, the frequency distribution was given by $p_i^{(m)} = c_m/i^\alpha$, $i = 1, 2, \ldots, m$, with $\alpha = 0.7$. Recall that t and m are related by $t = \delta\, m/c$ in the result for $\alpha < 1$ in Sect. 2.3, and two cases of $m = 2\,t$ and $m = 10\,t$ with $t = 1{,}000$ were executed; that is, $\delta/c = 1/2$ and $\delta/c = 1/10$, respectively. The value of normalization constant c_m was calculated as ${c_m}^{-1} = \sum_{i=1}^{m} 1/i^\alpha$ and the empirical size index distribution $S_k^{(m)}(t)/\overline{S}_1^{(m)}(t)$, $k = 1, 2, \ldots$, was estimated by simulation. To evaluate the limiting size index distribution $\Psi_k(\alpha, \delta)$, $k = 1, 2, \ldots$, given on the right-hand side of (12) in Corollary 1, on the other hand, we have to determine the value of constant δ, which is equivalent to determining the value of c in our example. Although we have discussed that $c = 1 - \alpha$ in the argument after Assumption 2, the problem is that $c_m \neq (1 - \alpha)/m^{1-\alpha}$ for finite m. Thus, the value of c was determined by the relation $c_m = c/m^{1-\alpha}$ and that of δ was determined by $\delta = c\,t/m = c_m\, t/m^\alpha$. As in the preceding example, we extend the range of the function $\Psi_k(\alpha, \delta)$, $k = 1, 2, \ldots$, on the right-hand side of (12) to $\mathbb{R}_+$, and applying the values of δ determined above, the curves of the function,

$$\Psi(x; \alpha, \delta) = \frac{\delta^{1/\alpha}\, \Gamma(x - 1/\alpha,\ \delta)}{\alpha\, \Gamma(x + 1)\, \left[\delta^{1/\alpha}\, \Gamma(1 - 1/\alpha,\ \delta) + 1 - e^{-\delta}\right]}, \quad x \geq 0, \tag{34}$$

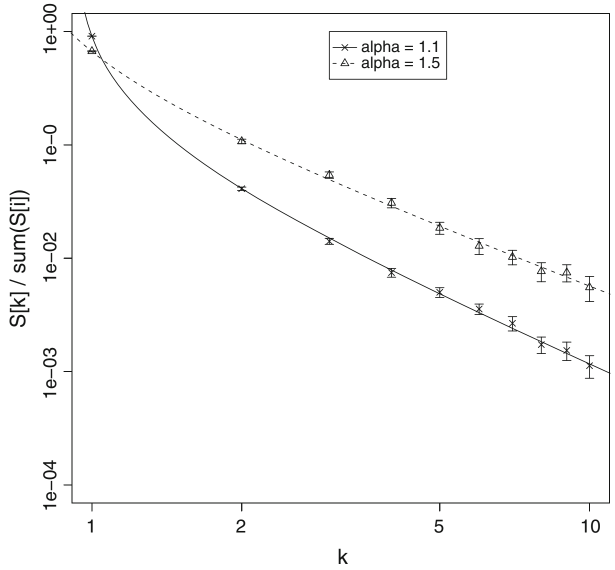

Fig. 1 Experimental results for Example 1 (Independent throw model with $\alpha > 1$)

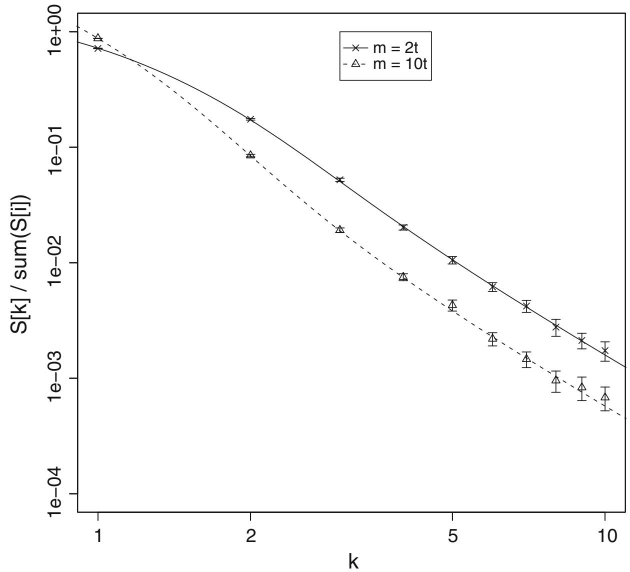

Fig. 2 Experimental results for Example 2 (Independent throw model with $\alpha = 0.7$)

are evaluated. The results are compared in Fig. 2, where we can also see good agreement for both cases of $m = 2t$ and $m = 10t$. Furthermore, from $\delta\big|_{m=10t} = c_{10t}\, t^{1-\alpha}/10^{\alpha} < c_{2t}\, t^{1-\alpha}/2^{\alpha} = \delta\big|_{m=2t}$, we can see in the experimental results that the size index distribution tends to degenerate as in (13) as the value of δ decreases.

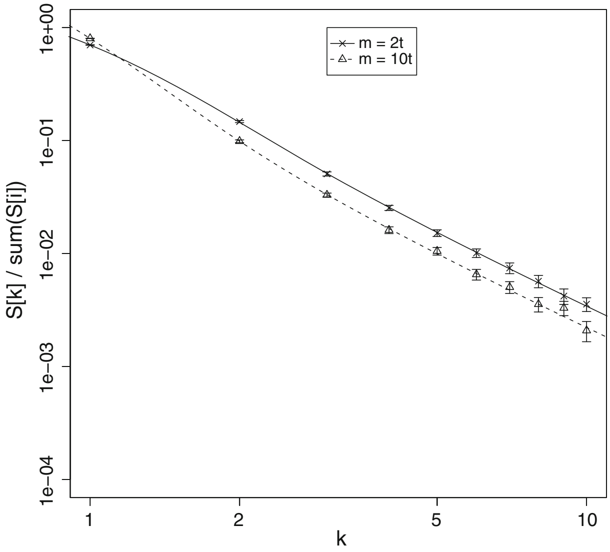

Fig. 3 Experimental results for Example 3 (Independent throw model with $\alpha = 1$)

Example 3 (Case of $\alpha = 1$) The third example is almost the same as Example 2 but the case of $\alpha = 1$ is examined. As in the preceding example, two cases of $m = 2t$ and $m = 10t$ were executed with $t = 1,000$. The value of the normalization constant c_m was calculated as $c_m{}^{-1} = \sum_{i=1}^{m} 1/i$ for the simulation runs. To evaluate the theoretical formula (18) in Corollary 2, the value of δ was then determined according to the relation $t = (\delta\, m \ln m)/c$; that is, the value of c was given by $c_m = c/\ln m$ and the value of δ was determined by $\delta = c\,t/(m \ln m) = c_m\, t/m$. The experimental results are given in Fig. 3, where the curves of function (34) with $\alpha = 1$ are plotted to compare them with the simulation estimates. Again, we can see good agreement for both cases of $m = 2t$ and $m = 10t$.

4.2 Dependent throw model

We here examine the dependent throw model. To achieve dependence, we consider the random intensity process $\{\boldsymbol{q}(t)\}_{t\ge0}$ governed by a two-state Markov chain $\{M(t)\}_{t\ge0}$ on $\{0, 1\}$ such that

$$q_i^{(m)}(t) = \begin{cases} c_m^{(\text{even})}/i^{\alpha}\, 1_{\{M(t)=0\}}, & i \text{ is even,} \\ c_m^{(\text{odd})}/i^{\alpha}\, 1_{\{M(t)=1\}}, & i \text{ is odd,} \end{cases} \qquad t \ge 0; \tag{35}$$

that is, when the Markov chain is in state 0, only bins with even indices are chosen and, when it is in state 1, only bins with odd indices are chosen. In the experiments discussed below, the parameter of sojourn times at each state was set at $1/5$; that is, the Markov chain stays at a state during a time according to the exponential distribution

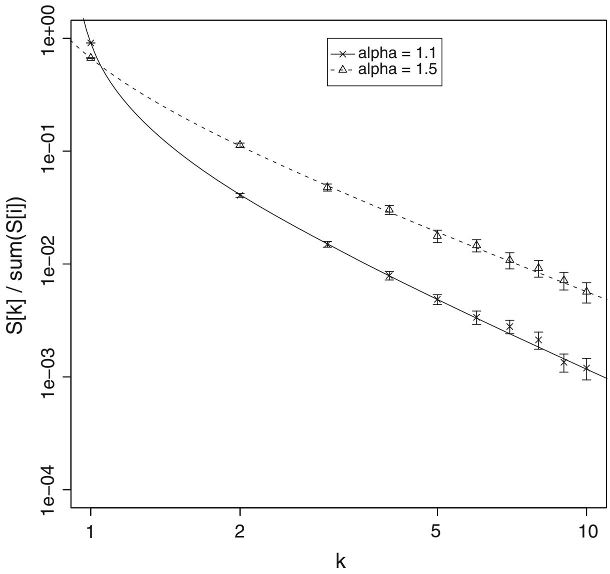

Fig. 4 Experimental results for Example 4 (Dependent throw model with $\alpha > 1$)

with mean 5 and moves to the other state. The stationary distribution of this Markov chain is given as (1/2, 1/2).

Example 4 (Case of $\alpha > 1$) As in Example 1, two cases of $\alpha = 1.1$ and $\alpha = 1.5$ were executed, where the values of $c^{(\text{even})}$ and $c^{(\text{odd})}$ in (35) were respectively calculated as

$$c^{(\text{even})\,-1} = \sum_{i=1}^{i_0} (2\,i)^{-\alpha} + \frac{(2\,i_0)^{-\alpha+1}}{2\,(\alpha-1)},$$
$$c^{(\text{odd})\,-1} = \sum_{i=1}^{i_0} (2\,i-1)^{-\alpha} + \frac{(2\,i_0-1)^{-\alpha+1}}{2\,(\alpha-1)},$$

with $i_0 = 10^7$. The experimental results are given in Fig. 4, where the solid line and the dashed line are the same as those in Fig. 1 since the right-hand side of (33) is irrelevant to the values of $c^{(\text{even})}$ and $c^{(\text{odd})}$. Even in the dependent throw model, we can see the same feature as in the independent throw model; that is, the theoretical results agree with those from the simulation experiment very well.

Example 5 (Case of $\alpha < 1$) As in Example 2, the power of the Zipf-type distribution was set at $\alpha = 0.7$, and two cases of $m = 2\,t$ and $m = 10\,t$ were executed for $t = 1{,}000$. The values of $c_m^{(\text{even})}$ and $c_m^{(\text{odd})}$ in (35) were respectively calculated as $c_m^{(\text{even})\,-1} = \sum_{i=1}^{m/2} 1/(2\,i)^{\alpha}$ and $c_m^{(\text{odd})\,-1} = \sum_{i=1}^{m/2} 1/(2\,i-1)^{\alpha}$. To evaluate the theoretical formula (34), the value of constant δ was determined as follows. Since the

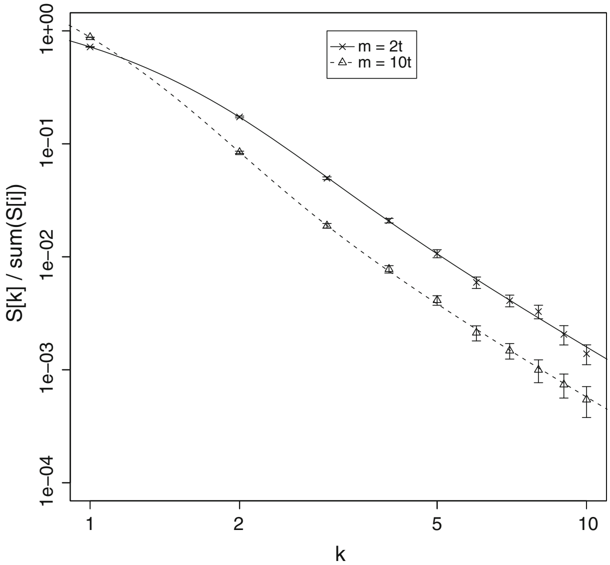

Fig. 5 Experimental results for Example 5 (Dependent throw model with $\alpha = 0.7$)

stationary distribution of the Markov chain $\{M(t)\}_{t\geq 0}$ is given as $(1/2, 1/2)$, we have

$$p_i^{(m)} = \begin{cases} (1/2)\, c_m^{(\text{even})}/\, i^{\alpha}, & i \text{ is even}, \\ (1/2)\, c_m^{(\text{odd})}/\, i^{\alpha}, & i \text{ is odd}. \end{cases}$$

The value of c_m was set at the middle of $c_m^{(\text{even})}/2$ and $c_m^{(\text{odd})}/2$; that is, $c_m = (c_m^{(\text{even})} + c_m^{(\text{odd})})/4$, and the value of δ was then determined by $\delta = c_m\, t/m^{\alpha}$ as in Example 2. Applying the values of δ determined as such, the curves of the function on the right-hand side of (34) are plotted in Fig. 5, where the simulation estimates are also plotted. Even in this case, we can see good agreement for both cases of $m = 2\,t$ and $m = 10\,t$.

Example 6 (Case of $\alpha = 1$) Finally, we examine the dependent throw model with $\alpha = 1$. The value of c_m was determined as in the preceding example with $\alpha = 1$ and the value of δ was set at $\delta = c_m\, t/m$ as in Example 3. The experimental results are given in Fig. 6, where we can also see good agreement.

A Properties of $\Phi_k(\alpha, \delta)$ and $\Psi_k(\alpha, \delta)$ as $\delta \downarrow 0$

We here verify the assertions made in Remarks 1 and 2 in Sect. 2. For $\alpha > 0$ and $\delta > 0$, $\Phi_k(\alpha, \delta)$, $k = 0, 1, 2, \ldots$, and $\Psi_k(\alpha, \delta)$, $k = 1, 2, \ldots$, are given by

$$\Phi_k(\alpha, \delta) = \frac{\delta^{1/\alpha}}{\alpha\, k!}\, \Gamma\left(k - \frac{1}{\alpha},\, \delta\right), \quad k = 0, 1, 2, \ldots, \tag{36}$$

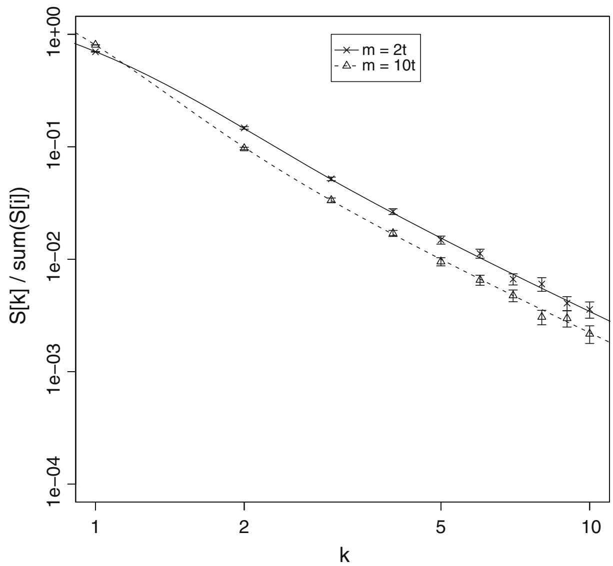

Fig. 6 Experimental results for Example 6 (Dependent throw model with $\alpha = 1$)

$$\Psi_k(\alpha, \delta) = \frac{\Phi_k(\alpha, \delta)}{\sum_{l=1}^{\infty} \Phi_l(\alpha, \delta)} = \frac{\Gamma(k-1/\alpha, \delta)}{\alpha\, k!\, \left[\Gamma(1-1/\alpha, \delta) + \delta^{-1/\alpha}(1-e^{-\delta})\right]}, \quad k = 1, 2, \ldots. \tag{37}$$

Lemma 7 (i) *For* $\alpha > 0$, $\Phi_k(\alpha, \delta)$, $k = 0, 1, 2, \ldots$, *in* (36) *degenerates as in* (11) *as* $\delta \downarrow 0$.

(ii) *When* $\alpha > 1$, $\Psi_k(\alpha, \delta)$,$k = 1, 2, \ldots$, *in* (37) *reduces to the KRS distribution; that is, the right-hand side of* (2)*, as* $\delta \downarrow 0$. *When* $\alpha \in (0, 1]$, *it degenerates as in* (13) *as* $\delta \downarrow 0$.

Proof We first show (i). Since $\Gamma(x, y) \to \Gamma(x) < \infty$ as $y \downarrow 0$ for $x > 0$, it is clear from (36) that $\Phi_k(\alpha, \delta) \to 0$ as $\delta \downarrow 0$ for $k > 1/\alpha$. For $k \leq 1/\alpha$, since $\Gamma(x, y) \to \infty$ as $y \downarrow 0$ for $x \leq 0$, applying de l'Hospital's rule, we have $\Phi_k(\alpha, \delta) \sim e^{-\delta}\delta^k/k!$ as $\delta \downarrow 0$, so that (11) is derived.

Next, we show (ii). Applying de l'Hospital's rule to the second term in the brackets of the denominator in (37), we have

$$\delta^{-1/\alpha}(1-e^{-\delta}) \sim \frac{\alpha\, e^{-\delta}}{\delta^{1/\alpha-1}} \to \begin{cases} 0, & \alpha > 1, \\ 1, & \alpha = 1, \\ +\infty, & 0 < \alpha < 1, \end{cases} \quad \text{as } \delta \downarrow 0. \tag{38}$$

Thus, since $\Gamma(x, y) \to \Gamma(x) < \infty$ as $y \downarrow 0$ for $x > 0$, the limit of $\Psi_k(\alpha, \delta)$ as $\delta \downarrow 0$ clearly reduces to the KRS distribution when $\alpha > 1$. When $\alpha = 1$, since $\Gamma(0, y) \to \infty$ as $y \downarrow 0$, it is also clear from (37) and (38) that (13) holds. Consider the case of $0 < \alpha < 1$. Note that $\Gamma(x, y) \to \Gamma(x) < \infty$ for $x > 0$ and $\Gamma(x, y) \to \infty$

for $x \leq 0$ as $y \downarrow 0$. Thus, when $k > 1/\alpha$, clearly $\Psi_k(\alpha, \delta) \to 0$ as $\delta \downarrow 0$ by (37) and (38). To consider the case of $k \leq 1/\alpha$, we deform the right-hand side of (37) as

$$\Psi_k(\alpha, \delta) = \frac{1}{\alpha\, k!} \frac{\Gamma(k - 1/\alpha, \delta)}{\Gamma(1 - 1/\alpha, \delta)} \left[1 + \frac{\delta^{-1/\alpha}\,(1 - e^{-\delta})}{\Gamma(1 - 1/\alpha, \delta)} \right]^{-1}.$$

Here, applying de l'Hospital's rule to the second term in the brackets, we have

$$\frac{\delta^{-1/\alpha}\,(1 - e^{-\delta})}{\Gamma(1 - 1/\alpha, \delta)} \sim \frac{(1/\alpha)\,\delta^{-1}\,(1 - e^{-\delta}) - e^{-\delta}}{e^{-\delta}} \to \frac{1}{\alpha} - 1 \quad \text{as } \delta \downarrow 0,$$

where we use (38). Furthermore, applying de l'Hospital's rule again, we have $\Gamma(k - 1/\alpha, \delta)/\,\Gamma(1-1/\alpha, \delta) \sim \delta^{k-1}$ as $\delta \downarrow 0$. Hence, we eventually have (13) for $0 < \alpha < 1$. □

Acknowledgments The authors are grateful to Shigeru Mase for his profuse encouragement and useful comments, as well as for his advice on the use of the R software. The first author wishes to thank Takafumi Kanamori for the valuable discussions. Thanks are also due to Hirotaka Yamashita for certain suggestions on the proof in the Appendix. Finally, the authors would like to thank the anonymous reviewers for their helpful suggestions and comments that have greatly improved this work.

References

Baayen, R. H. (2001). *Word frequency distributions*. Dordrecht: Kluwer.

Davis, P. J. (1972). Gamma function and related functions. In M. Abramowitz, I. A. Stegun (Eds.), *Handbook of mathematical functions with formulas, graphs, and mathematical tables, 9th printing* (Chap. 6, pp. 253–293). New York: Dover.

Evert, S. (2004). A simple LNRE model for random character sequences. *Proceedings of the 7èmes Journées Internationales d'Analyse Statistique des Données Textuelles (JADT2004)*, 411–422.

Gnedin, A., Hansen, B., Pitman, J. (2007). Notes on the occupancy problem with infinitely many boxes: General asymptotics and power laws. *Probability Surveys, 4*, 146–171.

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

Ivanov, V. A., Ivchenko, G. I., Medvedev, Yu. I. (1985). Discrete problems in probability theory. *Journal of Mathematical Sciences (New York), 31*, 2759–2795.

Karlin, S. (1967). Central limit theorems for certain infinite urn schemes. *Journal of Mathematics and Mechanics, 17*, 373–401.

Khmaladze, E. V. (1987). The statistical analysis of a large number of rare events. Report MS-R8804, Centre for Mathematics and Computer Science, CWI, Amsterdam, The Netherlands.

Mitzenmacher, M., Upfal, E. (2005). *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge: Cambridge University Press.

Müller, A., Stoyan, D. (2002). *Comparison methods for stochastic models and risks*. Chichester: Wiley.

R Development Core Team (2007). R: A language and environment for statistical computing. http://www.R-project.org.

Rouault, A. (1978). Lois de Zipf et sources markoviennes. *Annales de l'Institut Henri Poincaré: Probabilités et Statistiques, 14*, 169–188.

Shaked, M., Shanthikumar, J. G. (2007). *Stochastic orders*. New York: Springer.

Sibuya, M. (1979). Generalized hypergeometric, digamma and trigamma distributions. *Annals of the Institute of Statistical Mathematics, 31*, 373–390.

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