S©MR ISSN 1813-3304

СИБИРСКИЕ ЭЛЕКТРОННЫЕ
МАТЕМАТИЧЕСКИЕ ИЗВЕСТИЯ

Siberian Electronic Mathematical Reports
http://semr.math.nsc.ru

Том 18, стр. 1035–1045 (2021) УДК 519.214.5
DOI 10.33048/semi.2021.18.079 MSC 60F17

ON THE ACCURACY OF THE POISSONISATION IN THE INFINITE OCCUPANCY SCHEME

M. CHEBUNIN

Abstract. We obtain asymptotic accuracy of the poissonisation in the infinite occupancy scheme. Some of the results are obtained for integer-valued random variables having a regularly varying distribution.

Keywords: infinite urn/cell scheme, asymptotic upper bounds, poissonisation.

1. Introduction

We consider a model with n balls and infinitely many cells ("urns") numbered $1, 2, \ldots$. Ball $j = 1, 2, \ldots, n$ is randomly thrown to cell X_j, $\mathbb{P}(X_j = i) = p_i > 0$, $\sum\limits_{i=1}^{\infty} p_i = 1$, independently of everything else. Denote by $J_i(n) = \sum\limits_{j=1}^{n} \mathbb{I}(X_j = i)$ the total number of balls in cell i. Let

$$R^*_{n,k} = \sum_{i=1}^{\infty} \mathbb{I}(J_i(n) \geq k) \tag{1}$$

be the number of cells containing at least $k \geq 1$ balls,

$$R_{n,k} = R^*_{n,k} - R^*_{n,k+1} = \sum_{i=1}^{\infty} \mathbb{I}(J_i(n) = k) \tag{2}$$

the number of cells with exactly k balls, and assume $p_1 \geq p_2 \geq \ldots$.

Chebunin, M., On the accuracy of the poissonisation in the infinite occupancy scheme.

The work is supported by Mathematical Center in Akademgorodok under agreement No. 075-15-2019-1675 with the Ministry of Science and Higher Education of the Russian Federation.

Received September, 1, 2021, published October, 12, 2021.

Karlin (1967) has obtained pioneering results in the study of this model. We recall here a number of his results. It seems that he was the first who introduced the "poissonisation"procedure in this content. Namely, instead of fixed-size sampling he considered samples of random size $P(n)$, where $\{P(t),\ t \geq 0\}$ is a Poisson process with intensity one that does not depend on the procedure of assigning cells to balls. According to the well-known splitting property of Poisson flows, random processes $\{J_i(P(t)) \stackrel{def}{=} P_i(t),\ t \geq 0\}$ are Poisson with intensities $p_i,\ i = 1, 2, \ldots,$ and mutually independent for different i. From (1),

$$R^*_{P(t),k} = \sum_{i=1}^{\infty} \mathbb{I}(P_i(t) \geq k) \ \text{ and } \ R_{P(t),k} = \sum_{i=1}^{\infty} \mathbb{I}(P_i(t) = k).$$

Let $\alpha(x) = \max\{j \ : \ p_j \geq 1/x\}$ and assume the function $\alpha(x)$ to be regularly varying at infinity,

$$\alpha(x) = x^{\theta} L(x) \ \text{ with } \ \theta \in [0, 1], \tag{3}$$

where $L(x)$ is a function slowly varying at infinity. Clearly, $L(t) \to 0$ as $t \to \infty$, if $\theta = 1$. Lemma 4 (if $\theta = 1$) of Karlin showed that function

$$L^*(t) \stackrel{def}{=} \int_0^{\infty} \frac{e^{-1/y}}{y} L(ty)dy \to 0 \ \text{ as } \ t \to \infty$$

is slowly varying, too. Let, for $k \geq 1$, $n \geq 1$, $t > 0$

$$Y^*_{n,k} = R^*_{n,k} - \mathbb{E}R^*_{n,k}, \quad Y_{n,k} = R_{n,k} - \mathbb{E}R_{n,k},$$
$$Z^*_k(t) = R^*_{P(t),k} - \mathbb{E}R^*_{P(t),k}, \quad Z_k(t) = R_{P(t),k} - \mathbb{E}R_{P(t),k},$$
$$\Phi^*_k(t) = \mathbb{E}R^*_{P(t),k},\ \Phi_k(t) = \mathbb{E}R_{P(t),k},\ V^*_k(t) = \mathbb{V}arR^*_{P(t),k},\ V_k(t) = \mathbb{V}arR_{P(t),k},$$

and let $R_n \stackrel{def}{=} R^*_{n,1} = \sum_{k \geq 1} R_{n,k}$ be the number of non-empty cells ($\Phi(t) \stackrel{def}{=} \Phi^*_1(t)$, $V(t) \stackrel{def}{=} V^*_1(t)$). Karlin has established a number of asymptotic properties of random variables R_n as $n \to \infty$ under condition (3), including the Strong Law of Large Numbers (SLLN) and the asymptotic normality in the range $\theta \in (0, 1]$, and also the asymptotic normality of random vector $(R_{n,1}, ..., R_{n,k})$, $k \geq 1$ when $\theta \in (0, 1)$. The proof of normality was based on the following convergences: as $n \to \infty$

$$\mathbb{E}R_n - \Phi(n) \to 0 \tag{4}$$

and under condition (3), for any fixed $c_0 > 0$, $\theta \in (0, 1]$ and c_1

$$\sup_{|c| \leq c_0} \frac{|\mathbb{E}R_{[n+c\sqrt{n}]} - \mathbb{E}R_n|}{\sqrt{V(n)}} \to 0, \tag{5}$$

$$\frac{R_{[n+c_1\sqrt{n}]} - R_n}{\sqrt{V(n)}} \xrightarrow{\text{p}} 0. \tag{6}$$

Dutko (1989) has proved the asymptotic normality of R_n under a weaker assumption. Namely, he replaced regular condition (3) by the following:

$$V(n) \to \infty \ \text{ as } \ n \to \infty. \tag{7}$$

For the regularly varying tails, condition (3) holds for any positive θ and may also hold for $\theta = 0$ in a particular case. Dutko did not assume condition (3) in his proofs of (5) and (6).

Gnedin, Hansen and Pitman (2007) have studied sufficient conditions for (7), found rate of convergence in (4) and provided an overview on the topic.

Hwang and Janson (2008) have proved local limit theorems for a finite and infinite number of cells.

Barbour and Gnedin (2009) have proved asymptotic normality of random vector $(R_{n,1},\dots,R_{n,k})$ for $k \geq 1$ under the condition $V_i(n) \to \infty$ as $n \to \infty$, for any $i = 1,\dots,k$. Note that it is sufficient to have $V_k(n) \to \infty$ or $\Phi_k(n) \to \infty$ (see Lemma 5). They have obtained (in their Lemma 2.1) an upper bound for the total variation distance between vectors $(R_{n,1},\dots,R_{n,k})$ and $(R_{P(n),1},\dots,R_{P(n),k})$, and also showed that the covariance matrices converge if and only if condition (3) holds.

Barbour (2009) has proved theorems on approximation of the number of cells with k balls by translated Poisson distribution in the total variation distance.

Chebunin and Kovalevskii (2016) have proved the Functional Central Limit Theorem for random vector $(R^*_{n,1},...,R^*_{n,k})$, $\theta \in (0,1)$, $k \geq 1$. Their proof is based on the convergence

$$\sup_{0\leq t\leq 1} \frac{|R_{P(nt)} - R_{[nt]}|}{\sqrt{V(n)}} \xrightarrow{\text{p}} 0 \quad \text{as } n \to \infty.$$

Zakrevskaya and Kovalevskii (2001) have proposed an implicit estimator of parameter θ based on R_n for one-parametric family and proved its consistency.

Chebunin (2014) has proposed explicit estimators of the parameter based on R_n for a broader class of distributions and proved their consistency.

Chebunin and Kovalevskii (2019a, 2019b) have proposed asymptotically normal estimators of the parameter based on R_n, and statistical test for correspondence to the Zipf's law.

In this paper, we analyse accuracy of a.s. approximation of $R^*_{n,k}$ by $R^*_{P(n),k}$ when n grows, for any fixed $k \geq 1$.

Let $t^{\pm} = t \pm 2\sqrt{t \ln\ln t}$ for $t > e$, and for $k \geq 1$

$$\Delta^*_{n,k} \stackrel{def}{=} R^*_{P(n^+),k} - R^*_{P(n^-),k}, \quad a^*_{n,k} = \mathbb{E}\Delta^*_{n,k}, \quad \sigma^{*2}_{n,k} = \mathbb{V}ar\Delta^*_{n,k}, \quad \beta^*_{n,k} = \sigma^{*2}_{n,k}/\ln n.$$

Denote $f(x) = x - (1+x)\ln(1+x)$ the non-increasing negative function for $x \geq 0$. Let $\delta_0 = \delta_0(\delta)$ be the solution of the equation $f(x) = -\frac{1}{\delta}$, for some fixed $\delta > 0$.

Theorem 1. *For any $k \geq 1$*

$$\mathbb{P}\left(\limsup_{n\to\infty} b^*_{n,k}|R^*_{n,k} - R^*_{P(n),k}| \leq 1\right) = 1, \quad \mathbb{P}\left(\limsup_{n\to\infty} b_{n,k}|R_{n,k} - R_{P(n),k}| \leq 1\right) = 1,$$

where $b_{n,k} = \min(b^*_{n,k}, b^*_{n,k+1})/2$ and

$$b^*_{n,k} = \begin{cases} \min\{1/l^*_{n,k}, 1\}, & \text{if } a^*_{n,k} \to 0 \text{ as } n \to \infty; \\ \min\{1/(2l^*_{n,k}), 1/(2a^*_{n,k})\}, & \text{otherwise.} \end{cases} \tag{8}$$

Here

1) $l^*_{n,k} = \sqrt{2\sigma^{*2}_{n,k}\ln n}$ *if* $\beta^*_{n,k} \to \infty$ *as* $n \to \infty$

2) $l^*_{n,k} = \delta_0\sigma^{*2}_{n,k}$ *if* $\liminf\limits_{n\to\infty} \beta^*_{n,k} = \delta > 0$

3) $l^*_{n,k} = (e-1)\max(\sigma^{*2}_{n,k}, \ln n)$ *if* $\limsup\limits_{n\to\infty} \beta^*_{n,k} \neq \liminf\limits_{n\to\infty} \beta^*_{n,k} = 0$

4) $l^*_{n,k} = \gamma^*_{n,k}\ln n$ *for any* $\gamma^*_{n,k} \sim -1/\ln\beta^*_{n,k}$ *if* $\beta^*_{n,k} \to 0$ *as* $n \to \infty$.

Since the variance of an indicator random variable is not bigger than its expectation, then $\beta^*_{n,k}\ln n = \sigma^{*2}_{n,k} \le a^*_{n,k}$, and if $a^*_{n,k} \to 0$ as $n \to \infty$ then $\beta^*_{n,k} \to 0$ as $n \to \infty$.

Corollary 1. *If $\mathbb{E}X_1^{1+\varepsilon} < \infty$ for some $\varepsilon > 0$ then for $k \ge 1$*

$$\mathbb{P}\left(\limsup_{n\to\infty} |R^*_{n,k} - R^*_{P(n),k}| \le \frac{2(2+\varepsilon)}{\varepsilon}\right) = 1,$$

$$\mathbb{P}\left(\limsup_{n\to\infty} |R_{n,k} - R_{P(n),k}| \le \frac{4(2+\varepsilon)}{\varepsilon}\right) = 1.$$

Moreover, if $\mathbb{E}X_1^M < \infty$ for any $M \ge 1$ then for $k \ge 1$

$$\mathbb{P}\left(\limsup_{n\to\infty} |R^*_{n,k} - R^*_{P(n),k}| \le 2\right) = 1, \quad \mathbb{P}\left(\limsup_{n\to\infty} |R_{n,k} - R_{P(n),k}| \le 4\right) = 1.$$

Note that if condition (3) holds with $\theta < 1/2$, then there exists $\varepsilon > 0$ such that $\mathbb{E}X_1^{1+\varepsilon} < \infty$ (for more details see the next corollary).

Corollary 2. *Under condition (3) there exists a constant $C = C(L) > 0$ such that*

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{|R^*_{n,k} - R^*_{P(n),k}|}{\phi_k(n)} \le 1\right) = 1, \quad \mathbb{P}\left(\limsup_{n\to\infty} \frac{|R_{n,k} - R_{P(n),k}|}{2\phi_k(n)} \le 1\right) = 1,$$

where

$$\phi_k(n) = \begin{cases} 8\sqrt{n\ln\ln n}L^*(n), & \theta = 1,\ k = 1; \\ \frac{8\theta\Gamma(k-\theta)}{(k-1)!} n^{\theta-\frac{1}{2}}\sqrt{\ln\ln n}L(n), & \theta \in (1/2,1) \vee\ (\theta = 1,\ k \ge 2); \\ C\max(\sqrt{\ln\ln n}L(n), \ln n), & \theta = 1/2; \\ 2/(1-2\theta), & \theta < 1/2. \end{cases} \tag{9}$$

Note that for $\theta = 1$ sequence $\phi_k(n)$ is a little bit smaller than what we could expect to appear in the Law of the Iterated Logarithm (LIL) for $R_{n,k}$. For $\theta \in [1/2, 1)$ sequence $\phi_k(n)$ is smaller than the normalization in the CLT. For $\theta < 1/2$ sequence $\phi_k(n)$ don't depend on n. As a corollary, we obtain asymptotic upper bounds for the absolute values of $Y^*_{n,k}$.

Remark 1. *As it follows from Lemma 1 in Gnedin, Hansen, Pitman (2007), $\mathbb{E}R_n - \Phi(n) \to 0$, $\mathbb{E}R_{n,k} - \Phi_k(n) \to 0$ as $n \to \infty$. Then $\mathbb{E}R^*_{n,k} - \Phi^*_k(n) \to 0$ too, since $R^*_{n,k} = R_n - R_{n,1} - \ldots - R_{n,k-1}$. To establish the LIL for non-random scheme of size $n = 1, 2, \ldots$, it suffices to prove the LIL for the poissonized scheme, with standard normalisation. We could not manage to prove the latter. However, we obtain a weaker result (see the next corollary) which may be viewed as an analogue of the LIL for arrays of random variables (see Sung (1996) and Hoffmann, Miao, Li, Xu (2016) for further comments and background).*

Corollary 3. *Assume (3) to hold. If $\frac{\Phi_{k_0}(n)}{\ln n} \to \infty$ as $n \to \infty$, for some $k_0 \ge 1$, then, for any $k \le k_0$,*

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{|Y^*_{n,k}|}{\sqrt{2V^*_k(n)\ln n}} \le 1\right) = 1, \quad \mathbb{P}\left(\limsup_{n\to\infty} \frac{|Y_{n,k}|}{\sqrt{2V_k(n)\ln n}} \le 1\right) = 1.$$

Note that for $\theta > 0$ the assumption of Corollary 3 is held for all $k_0 \ge 1$ (this follows from the asymptotics of $\Phi_{k_0}(n)$).

The rest of the paper is organized as follows. In Sections 2 and 3 we formulate all the auxiliary results and prove Theorem 1 and corollaries. Appendix contains proofs of auxiliary results.

2. Proof of Theorem 1

For any $n \geq 1$, let random variables $\{\xi_{n,i}\}_{i\geq 1}$ in any row n be mutually independent with $\mathbb{E}\xi_{n,i} = 0$ for $i \geq 1$. Let $S_{n,N} = \sum\limits_{i=1}^{N} \xi_{n,i}$, $s_{n,N}^2 = \sum\limits_{i=1}^{N} \mathbb{E}\xi_{n,i}^2$ for $N \geq 1$, $S_n = S_{n,\infty}$, $s_n^2 = s_{n,\infty}^2 > 0$, and $h_n = s_n^2/\ln n$. We prove the following lemma for any dependence between rows.

Lemma 1. *Let $\mathbb{P}(\xi_{n,i} \leq c) = 1$ for all $n, i \geq 1$ and $s_n^2 < \infty$ for any fixed n. Then*

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{S_n}{l_n} \leq 1\right) = 1,$$

where
1) $l_n = \sqrt{2s_n^2 \ln n}$ if $h_n \to \infty$,
2) $l_n = \frac{c_0}{c} s_n^2$ if $\liminf\limits_{n\to\infty} h_n = \delta > 0$, where c_0 is the solution of the equation

$$f(x) = -\frac{c^2}{\delta},$$

3) $l_n = c_1 \max(s_n^2, \ln n)$ if $\limsup\limits_{n\to\infty} h_n \neq \liminf\limits_{n\to\infty} h_n = 0$, where c_1 is the solution of the equation $f(cx) = -c^2$,
4) $l_n = \gamma_n \ln n$ for any $\gamma_n \sim -c/\ln h_n$ if $h_n \to 0$.

Proof. By Borel-Cantelli lemma, it suffices to show that, for any $\eta > 1$,

$$\sum_{n=1}^{\infty} \mathbb{P}(S_n > \eta l_n) < \infty.$$

From identity $e^x = 1 + x + \psi(x)x^2$ for all $x \in \mathbb{R}$, where $\psi(x) = (e^x - 1 - x)/x^2$ be non-decreasing positive-valued function with $\psi(0) = 1/2$, we have for $t > 0$

$$\mathbb{E}e^{t\xi_{n,i}} = 1 + t^2\mathbb{E}(\xi_{n,i}^2\psi(t\xi_{n,i})) \leq 1 + t^2\psi(ct)\mathbb{E}\xi_{n,i}^2 \leq \exp\left\{t^2\psi(ct)\mathbb{E}\xi_{n,i}^2\right\}.$$

Since $\{\xi_{n,i}\}_{i=1}^{\infty}$ are mutually independent, we have, for any $N \geq 1$,

$$\mathbb{E}e^{tS_{n,N}} = \prod_{i=1}^{N} \mathbb{E}e^{t\xi_{n,i}} \leq \exp\left\{t^2\psi(ct)s_{n,N}^2\right\}. \tag{10}$$

Since, $s_n^2 < \infty$ for any fixed n then, by the Kolmogorov's two-series theorem, we have, that $S_{n,N} \to S_n$ as $N \to \infty$ almost surely. Using Fatou's lemma and letting $N \to \infty$ in (10), we get

$$\mathbb{E}e^{tS_n} = \mathbb{E}\left(\lim_{N\to\infty} e^{tS_{n,N}}\right) = \mathbb{E}\left(\liminf_{N\to\infty} e^{tS_{n,N}}\right) \leq \liminf_{N\to\infty} \mathbb{E}e^{tS_{n,N}}$$

$$\leq \liminf_{N\to\infty} \exp\left\{t^2\psi(ct)s_{n,N}^2\right\} = \lim_{N\to\infty} \exp\left\{t^2\psi(ct)s_{n,N}^2\right\} = \exp\left\{t^2\psi(ct)s_n^2\right\}.$$

Let $t = \frac{1}{c}\ln\left(1 + \frac{\eta c l_n}{s_n^2}\right)$ for any fixed $n \geq 1$. Then, by the Markov inequality

$$\mathbb{P}(S_n > \eta l_n) \leq \frac{\mathbb{E}e^{tS_n}}{e^{t\eta l_n}} \leq \exp\left\{\frac{\eta l_n}{c} - \frac{s_n^2 + \eta c l_n}{c^2}\ln\left(1 + \frac{\eta c l_n}{s_n^2}\right)\right\} = \exp\left\{\frac{s_n^2}{c^2} f\left(\frac{\eta c l_n}{s_n^2}\right)\right\}.$$

1) Notice that $f(x) \le -\frac{x^2}{2}(1-x)$ for $x > 0$ since $\ln(1+x) \ge x - \frac{x^2}{2}$ for $x \ge 0$. Since $l_n/s_n^2 = \sqrt{2/h_n} \to 0$ as $n \to \infty$,

$$\frac{s_n^2}{c^2} f\left(\frac{\eta c l_n}{s_n^2}\right) \le -\frac{\eta^2 l_n^2}{2s_n^2}\left(1 - \frac{\eta c l_n}{s_n^2}\right) = -\eta^2(1-o(1))\ln n.$$

2) Since $f(c_0) = -c^2/\delta$ and $f'(x) = -\ln(1+x) < 0$ for $x > 0$,

$$\frac{s_n^2}{c^2} f\left(\frac{\eta c l_n}{s_n^2}\right) = \frac{f(c_0\eta)}{c^2} s_n^2 \le \frac{\delta f(c_0\eta)}{c^2}(1+o(1))\ln n = -\eta'(1+o(1))\ln n, \text{ where } \eta' > 1.$$

3) Let $f_1(x) = f(x)/x$ for $x > 0$. Since $f_1(x) < 0$ and $f_1'(x) < 0$ for $x > 0$,

$$\frac{s_n^2}{c^2} f\left(\frac{\eta c l_n}{s_n^2}\right) = \frac{\eta l_n}{c} f_1\left(\frac{\eta c l_n}{s_n^2}\right) \le \frac{\eta l_n}{c} f_1(\eta c c_1) \le \frac{f(\eta c c_1)}{c^2}\ln n = -\eta''\ln n, \text{ where } \eta'' > 1.$$

4) Since $l_n/s_n^2 = \gamma_n/h_n \to \infty$ and $\gamma_n \to 0$,

$$\begin{aligned}\frac{s_n^2}{c^2} f\left(\frac{\eta c l_n}{s_n^2}\right) &\le \left(\frac{\eta\gamma_n}{c} - \frac{h_n + \eta c\gamma_n}{c^2}(\ln(\eta c\gamma_n) - \ln(h_n))\right)\ln n \\ &= \left(\frac{\eta}{c}\gamma_n \ln(h_n) - \frac{h_n}{c^2}\ln(\gamma_n) + o(1)\right)\ln n \\ = (-\eta(1+o(1)) + h_n \ln\ln(1/h_n)/c^2 + o(1))\ln n &= (-\eta + o(1))\ln n.\end{aligned}$$

□

Remark 2. *It is easy to see that if $h_n \le n^{-\kappa+o(1)}$ for some $\kappa > 0$ then*

$$-\frac{c}{\ln h_n} \le \frac{c}{(\kappa+o(1))\ln n}, \quad \textit{and } l_n \le c/\kappa.$$

Let $t_2 \stackrel{def}{=} t_2(n) \ge t_1 \stackrel{def}{=} t_1(n) \to \infty$ as $n \to \infty$, and for $k \ge 1$

$$\boldsymbol{a}_{k,n}^* \stackrel{def}{=} \boldsymbol{a}_{k,n}^*(t_1,t_2) = \mathbb{E}(R_{P(t_2),k}^* - R_{P(t_1),k}^*),$$

$$\boldsymbol{\sigma}_{n,k}^{*2} \stackrel{def}{=} \boldsymbol{\sigma}_{n,k}^{*2}(t_1,t_2) = \mathbb{V}ar(R_{P(t_2),k}^* - R_{P(t_1),k}^*),\ \boldsymbol{\beta}_{n,k}^* = \boldsymbol{\sigma}_{n,k}^{*2}/\ln n.$$

Lemma 2. *For any $k \ge 1$*

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{|Z_k^*(t_2) - Z_k^*(t_1)|}{\boldsymbol{l}_{n,k}^*} \le 1\right) = 1,$$

where
1) $\boldsymbol{l}_{n,k}^* = \sqrt{2\boldsymbol{\sigma}_{n,k}^{*2}\ln n}$ if $\boldsymbol{\beta}_{n,k}^* \to \infty$,
2) $\boldsymbol{l}_{n,k}^* = \delta_0\boldsymbol{\sigma}_{n,k}^{*2}$ if $\liminf_{n\to\infty} \boldsymbol{\beta}_{n,k}^* = \delta > 0$,
3) $\boldsymbol{l}_{n,k}^* = (e-1)\max(\boldsymbol{\sigma}_{n,k}^{*2}, \ln n)$ if $\limsup_{n\to\infty} \boldsymbol{\beta}_{n,k}^* \ne \liminf_{n\to\infty} \boldsymbol{\beta}_{n,k}^* = 0$,
4) $\boldsymbol{l}_{n,k}^* = \boldsymbol{\gamma}_{n,k}^* \ln n$ for any $\boldsymbol{\gamma}_{n,k}^* \sim -1/\ln\boldsymbol{\beta}_{n,k}^*$ if $\boldsymbol{\beta}_{n,k}^* \to 0$.

Proof. We use Lemma 1. Let

$$\xi_{n,i}^* = \pm(\mathbb{I}(P_i(t_2)) \ge k,\ P_i(t_1) < k) - \mathbb{P}(P_i(t_2)) \ge k,\ P_i(t_1) < k)) \le 1,$$

then $Z_k^*(t_2) - Z_k^*(t_1) = \sum_{i=1}^\infty \xi_{n,i}^* = S_n^*$ and $s_n^{*2} = \sum_{i=1}^\infty \mathbb{E}(\xi_{n,i}^*)^2$. As $R_{P(t),k}^* \le P(t)/k$ a.s., and the variance of an indicator random variable is not bigger than its expectation, we have $s_n^{*2} \le \boldsymbol{a}_{k,n}^* \le \mathbb{E}P(t_2)/k = t_2/k$.

□

Proof of Theorem 1. Recall basic properties of Poisson process. By the LIL for $P(t)$, for any $\eta > 1$, there exists $t_0 > 0$ such that

$$\mathbb{P}(\forall t \geq t_0,\ t - \eta\sqrt{2t\ln\ln t} < P(t) < t + \eta\sqrt{2t\ln\ln t}) = 1$$

Let $\eta = 2^{\frac{1}{4}}$, then there exists $t_0 > 0$ such that with probability 1 for any $t \geq t_0$

$$P(t^+) > t^+ - \eta\sqrt{2t^+\ln\ln t^+} > t + (2 - \eta^{\frac{3}{2}}\sqrt{2})\sqrt{t\ln\ln t} > t,$$

$$P(t^-) < t^- + \eta\sqrt{2t^-\ln\ln t^-} < t + (-2 + \eta\sqrt{2})\sqrt{t\ln\ln t} < t.$$

By monotonicity of $P(t)$, for any $\varepsilon \in (0,1)$, there exists an integer n_0 such that

$$\mathbb{P}(\forall n \geq n_0\ \exists \delta_n : |\delta_n| \leq 1,\ P(n + 2\delta_n\sqrt{n\ln\ln n}) = n) \stackrel{def}{=} \mathbb{P}(A(n_0)) \geq 1 - \frac{\varepsilon}{2} \tag{11}$$

Since $R^*_{n,k} = R^*_{P(n+\delta_n(n^+-n)),k}$ a.s., if $P(n + \delta_n(n^+ - n)) = n$, then conditionnaly on the event $A(n_0)$, we have, for any $n \geq n_0$ with probability one

$$|R^*_{n,k} - R^*_{P(n),k}| \leq \sup_{|\delta|\leq 1} |R^*_{P(n+\delta(n^+-n)),k} - R^*_{P(n),k}|$$

$$= \max(R^*_{P(n^+),k} - R^*_{P(n),k}, R^*_{P(n),k} - R^*_{P(n^-),k}) \leq R^*_{P(n^+),k} - R^*_{P(n^-),k} = \Delta^*_{n,k},$$

due to monotonicity of $P(t)$ in t and $R^*_{n,k}$ in n.

Clearly, the sequence $b^*_{n,k}$ in (8) satisfies condition $b^*_{n,k}a^*_{n,k} \leq 1/2$. So, by (8), (11) and Lemma 2, for any pair $\varepsilon > 0$, $\eta > 1$, there exists an integer n_0 such that for $n \geq n_0$

$$\mathbb{P}\left(\sup_{n\geq n_0} b^*_{n,k}|R^*_{n,k} - R^*_{P(n),k}| \geq \eta\right) \leq \mathbb{P}\left(\sup_{n\geq n_0} b^*_{n,k}|R^*_{n,k} - R^*_{P(n),k}| \geq \eta, A(n_0)\right) + \frac{\varepsilon}{2}$$

$$\leq \mathbb{P}\left(\sup_{n\geq n_0} b^*_{n,k}|\Delta^*_{n,k} \pm a^*_{n,k}| \geq \eta\right) + \frac{\varepsilon}{2}$$

$$\leq \mathbb{P}\left(\sup_{n\geq n_0} b^*_{n,k}|Z^*_k(n^+) - Z^*_k(n^-)| \geq \eta'\right) + \frac{\varepsilon}{2} \leq \varepsilon,$$

where

$$\eta' = \begin{cases} (\eta+1)/2, & \text{if } a^*_{n,k} \to 0 \text{ as } n \to \infty; \\ \eta/2, & \text{otherwise.} \end{cases}$$

The second assertion of the Theorem 1 follows directly from (2).

Theorem 1 is proved.

3. Proof of Corollaries

We need the following auxiliary results from Karlin (1967, Theorem 1, formulas (23), (26) and (37)). Assume (3) to hold, then as $t \to \infty$,

$$\Phi(t) \sim V(t) \sim \Phi_1(t) \sim V_1(t) \sim tL^*(t) \quad \text{if} \quad \theta = 1,$$

$$\Phi(t) \sim \Gamma(1-\theta)\alpha(t),\ V(t) \sim \Gamma(1-\theta)(2^\theta - 1)\alpha(t) \quad \text{if} \quad \theta \in (0,1),$$

$$\Phi(t) \sim \alpha(t),\quad V(t) \sim \alpha(2t) - \alpha(t) \quad \text{if} \quad \theta = 0,$$

$$\Phi_k(t) \sim \theta\frac{\Gamma(k-\theta)}{k!}\alpha(t),\quad V_k(t) \sim \frac{\theta}{k!}\left[\Gamma(k-\theta) - \frac{\Gamma(2k-\theta)}{2^{2k-\theta}k!}\right]\alpha(t)$$

if either $\theta \in (0,1),\ k \geq 1$ or $\theta = 1,\ k \geq 2$.

The proofs of the following lemmas may be found in Appendix.

Lemma 3. *Assume (3) to hold. For $k \geq 2$ and as $t \to \infty$*

$$\Phi_k^*(t) \sim \frac{\Gamma(k-\theta)}{\Gamma(k)}\alpha(t), \quad \theta \in [0,1].$$

If $\theta \in (0,1]$, then also

$$V_k^*(t) \sim \frac{\alpha(t)}{\Gamma(k)}\left(\sum_{i=0}^{k-1} 2^{\theta+1-k-i}\frac{\Gamma(k+i-\theta)}{\Gamma(i+1)} - \Gamma(k-\theta)\right).$$

Lemma 4. *Let $t_2 \geq t_1 > 0$ then, for any $k \geq 1$*

$$\Phi_k^*(t_2) - \Phi_k^*(t_1) \leq k\left(\frac{t_2}{t_1}\right)^{k-1}\frac{t_2-t_1}{t_1}\Phi_k(t_1).$$

Lemma 5. *Let d_n be a sequence of positive constants such that, $d_{[cn]}/d_n > \varepsilon(c) > 0$ for any $c > 0$, $n \geq n_0$. Then the conditions $\min\limits_{1\leq k\leq k_0} \min(V_k^*(n), V_k(n))/d_n \to \infty$ and $\Phi_{k_0}(n)/d_n \to \infty$ as $n \to \infty$ are equivalent.*

Proof of Corollary 1. Let $\mathbb{E}X_1^{1+\varepsilon} < \infty$ for some $\varepsilon > 0$, then $p_i = o(i^{-2-\varepsilon})$ as $i \to \infty$. Hence there exists a constant $c > 0$ such that $p_i \leq ci^{-2-\varepsilon}$ for $i \geq 1$, then $\alpha(x) \leq \max\{i : ci^{-2-\varepsilon} \geq 1/x\} \leq (cx)^{1/(2+\varepsilon)}$ and

$$\Phi(t) = \int_0^\infty \alpha(x)e^{-\frac{t}{x}}\frac{t}{x^2}dx \leq (ct)^{\frac{1}{2+\varepsilon}}\int_0^\infty y^{\frac{1}{2+\varepsilon}-2}e^{-\frac{1}{y}}dy = (ct)^{\frac{1}{2+\varepsilon}}\Gamma\left(1-\frac{1}{2+\varepsilon}\right).$$

Since $\Phi_k(t) \leq \Phi(t)$ for all $k \geq 1$, then from Lemma 4 we have

$$\beta_{n,k}^* \leq \frac{a_{n,k}^*}{\ln n} \leq \frac{k}{\ln n}\left(\frac{n^+}{n^-}\right)^{k-1}\frac{4\sqrt{n\ln\ln n}}{n^-}\Phi(n^-) \leq n^{\frac{1}{2+\varepsilon}-\frac{1}{2}+o(1)}.$$

Using remark 2 we get the required. If $\mathbb{E}X_1^M < \infty$ for all $M \geq 1$, then $\Phi(t) \leq t^{o(1)}$ and we again get the required using Remark 2.

Corollary 1 is proved.

Proof of Corollary 2. From Lemma 4 and asymptotics $\Phi_k(t)$ we have as $n \to \infty$

$$\beta_{n,k}^*\ln n = \sigma_{n,k}^{*2} \leq a_{n,k} \leq k\left(\frac{n^+}{n^-}\right)^{k-1}\frac{4\sqrt{n\ln\ln n}}{n^-}\Phi_k(n^-) \sim 4k\sqrt{\frac{\ln\ln n}{n}}\Phi_k(n).$$

We use Theorem 1. If $\theta < 1/2$, then from Remark 2 we get the required. If $\theta > 1/2$, then from asymptotics $\Phi_k(t)$ as $t \to \infty$ we get the required. Let $\theta = 1/2$, if $\sqrt{\ln\ln n}L(n)/\ln n \to \infty$ as $n \to \infty$, then $C(L) \leq 4\Gamma(k-1/2)/(k-1)!$, otherwise we can find $C(L) \leq \max(e-1, \delta_0)$ so that we get the required.

Corollary 2 is proved.

Proof of Corollary 3. We use Lemma 1. By Corollary 2, asymptotics for $V_k(t)$ and Lemma 3, it is enough to prove similar assertions for $Z_k^*(n)$ and $Z_k(n)$. Let, for $k \in \{1, ..., k_0\}$,

$$\xi_{n,i}^* = \pm(\mathbb{I}(J_i(P(n)) \geq k) - \mathbb{P}(J_i(P(n)) \geq k)),$$

$$\xi_{n,i} = \pm(\mathbb{I}(J_i(P(n)) = k) - \mathbb{P}(J_i(P(n)) = k)).$$

Then, for $n, i \geq 1$,

$$(s_n^*)^2 = V_k^*(n), \quad s_n^2 = V_k(n), \quad \text{and} \quad |\xi_{n,i}^*| \leq 1, \quad |\xi_{n,i}| \leq 1.$$

As $R_{P(n),k} \le R^*_{P(n),k} \le P(n)/k$ a.s., and the variance of an indicator random variable is not bigger than its expectation, we have

$$(s^*_n)^2 \le \Phi^*_k(n) \le \mathbb{E}P(n)/k = n/k.$$

Similarly, we get $s^2_n \le n/k$. Then by Lemma 5 the required result follows from Lemma 1.

Corollary 3 is proved.

Appendix

Proof of Lemma 3. Since $\alpha(x) = x^\theta L(x)$, for any integer $r \ge 2$ and as $t \to \infty$,

$$\int_0^\infty y^{-r-1}e^{-1/y}\alpha(ty)dy \sim \alpha(t)\int_0^\infty y^{\theta-r-1}e^{-1/y}dy = \alpha(t)\Gamma(r-\theta).$$

Denote by $f(x,t,k) = \sum\limits_{i=0}^{k-1} \frac{t^i}{x^i i!}e^{-t/x}$, then $f'_x(x,t,k) = \frac{t^k}{x^{k+1}(k-1)!}e^{-t/x}$.

Note that, for $k \ge 2,\ t > 0$,

$$\Phi^*_k(t) = \sum_{j=1}^\infty 1 - f(1/p_j,t,k) = \int_0^\infty (1-f(x,t,k))d\alpha(x)$$

$$= \int_0^\infty \alpha(x)f'_x(x,t,k)dx = \int_0^\infty \alpha(ty)f'_x(y,1,k)dy$$

$$= \frac{t^\theta}{(k-1)!}\int_0^\infty y^{\theta-k-1}e^{-1/y}L(ty)dy \sim \frac{\Gamma(k-\theta)}{\Gamma(k)}\alpha(t) \text{ as } t \to \infty.$$

The variance of $R^*_{P(t),k}$ for $\theta \in (0,1]$ may be found by

$$V^*_k(t) = \sum_{i=1}^\infty \mathbb{P}(P_i(t) \ge k)(1 - \mathbb{P}(P_i(t) \ge k))$$

$$= \sum_{i=1}^\infty \mathbb{P}(P_i(t) < k)(1 - \mathbb{P}(P_i(t) < k)) = \sum_{i=1}^\infty f(1/p_i,t,k)(1-f(1/p_i,t,k))$$

$$= \int_0^\infty f(x,t,k)(1-f(x,t,k))d\alpha(x)$$

We use integration by parts and then substitute $x = yt$:

$$V^*_k(t) = \int_0^\infty \alpha(x)(-f'_x(x,t,k) + 2f(x,t,k)f'_x(x,t,k))dx$$

$$= \int_0^\infty \alpha(yt)(2f(y,1,k)-1)f'_x(y,1,k)dy$$

$$\sim \frac{\alpha(t)}{\Gamma(k)}\left(\sum_{i=0}^{k-1} 2^{\theta+1-k-i}\frac{\Gamma(k+i-\theta)}{\Gamma(i+1)} - \Gamma(k-\theta)\right) \text{ as } t \to \infty.$$

Lemma 3 is proved.

Proof of Lemma 4. Let $t^* \in [t_1, t_2]$, then from the formula of finite Lagrange increments for the function $\Phi(t) = \sum_{i=1}^{\infty}(1 - e^{-tp_i})$

$$\Phi(t_2) - \Phi(t_1) = \frac{t_2 - t_1}{t_1} \sum_{i=1}^{\infty} t_1 p_i e^{-t^* p_i} \le \frac{t_2 - t_1}{t_1} \Phi_1(t_1).$$

Denote by $f(x,t,k) = \sum\limits_{i=0}^{k-1} \frac{t^i}{x^i i!} e^{-t/x}$, then $f'_t(x,t,k) = -\frac{t^{k-1}}{x^k (k-1)!} e^{-t/x}$. Similarly, from the formula of finite Lagrange increments for the function $\Phi_k^*(t) = \sum_{i=1}^{\infty}(1 - f(1/p_i, t, k))$, we obtain

$$\Phi_k^*(t_2) - \Phi_k^*(t_1) = k \frac{t_2 - t_1}{t_1} \left(\frac{t^*}{t_1}\right)^{k-1} \sum_{i=1}^{\infty} \frac{(t_1 p_i)^k}{k!} e^{-t^* p_i} \le k \left(\frac{t_2}{t_1}\right)^{k-1} \frac{t_2 - t_1}{t_1} \Phi_k(t_1).$$

Lemma 4 is proved.

Proof of Lemma 5. Note that

$$V_k^*(n) = \sum_{i=1}^{\infty} \mathbb{P}(P(np_i) < k)(1 - \mathbb{P}(P(np_i) < k))$$

$$\ge \sum_{i=1}^{\infty} \mathbb{P}(P(np_i) = 0)\mathbb{P}(P(np_i) = k) = \sum_{i=1}^{\infty} \frac{(np_i)^k}{k!} e^{-2np_i} = \frac{1}{2^k} \Phi_k(2n).$$

From Barbour and Gnedin (2009, formulas (4.1), (4.2), (4.4)), there exist positive constants c_k and $C_{k_0,k}$ such that $\Phi_k(n) > V_k(n) > c_k \Phi_k(n)$ and, for all $k < k_0$, the inequality $\Phi_k(n) \ge C_{k_0,k} \Phi_{k_0}(2n)$ holds. From Proposition 3.2 in Ben-Hamou, Boucheron, and Ohannessian (2017), we have $V_{k_0}^*(n) \le k_0 \Phi_{k_0}(n)$. Then

$$\Phi_{k_0}(n) > V_{k_0}(n) \ge \min_{1 \le k \le k_0} \min(V_k^*(n), V_k(n)) \ge \min_{1 \le k \le k_0} \min(\Phi_k(2n)/2^k, c_k \Phi_k(n))$$

$$\ge \min(2^{-k_0}, c_{k_0}, \min_{1 \le k < k_0} \min(C_{k_0,k}/2^k, c_k C_{k_0,k})) \cdot \min(\Phi_{k_0}(n), \Phi_{k_0}(2n), \Phi_{k_0}(4n)).$$

Since $d_{[cn]}/d_n \ge \varepsilon(c)$ for $n \ge n_0$,

$$\frac{\Phi_{k_0}([cn])}{d_n} = \frac{\Phi_{k_0}([cn])}{d_{[cn]}} \cdot \frac{d_{[cn]}}{d_n} \xrightarrow[n \to \infty]{} \infty \;\Leftrightarrow\; \frac{\Phi_{k_0}(n)}{d_n} \xrightarrow[n \to \infty]{} \infty.$$

Lemma 5 is proved.

Acknowledgements

The work is supported by Mathematical Center in Akademgorodok under agreement No. 075-15-2019-1675 with the Ministry of Science and Higher Education of the Russian Federation. The author would like to thank Sergey Foss and Artyom Kovalevskii for their constant attention to the work.

References

[1] A.D. Barbour, *Univariate approximations in the infinite occupancy scheme*, ALEA Lat. Am. J. Probab. Math. Stat., **6** (2009), 415–433. MR2576025

[2] A.D. Barbour, A.V. Gnedin, *Small counts in the infinite occupancy scheme*, Electron. J. Probab., **14** (2009), 365–384. MR2480545

[3] A. Ben-Hamou, S. Boucheron, M.I. Ohannessian, *Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications*, Bernoulli **23**:1 (2017), 249–287. MR3556773

[4] M.G. Chebunin, *Estimation of the parameters of probability models by the number of different elements in a sample*, Sib. Zh. Ind. Mat., **17**:3 (2014), 135–147. MR3364413
[5] M. Chebunin, A. Kovalevskii, *Functional central limit theorems for certain statistics in an infinite urn scheme*, Statist. Probab. Lett., **119** (2016), 344–348. MR3555307
[6] M. Chebunin, A. Kovalevskii, *Asymptotically normal estimators for Zipf's law*, Sankhya A, **81**:2 (2019), 482–492. MR4043483
[7] M. Chebunin, A. Kovalevskii, *A statistical test for the Zipf's law by deviations from the Heaps' law*, Sib. Électron. Mat. Izv., **16** (2019), 1822–1832. MR4050210
[8] M. Dutko, *Central limit theorems for infinite urn models*, Ann. Probab., **17**:3 (1989), 1255–1263. MR1009456
[9] A. Gnedin, B. Hansen, J. Pitman, *Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws*, Probab. Surv., **4** (2007), 146–171. MR2318403
[10] J. Hoffmann, Y. Miao, X.C. Li, S.F. Xu, *Kolmogorov type law of the logarithm for arrays*, J. Theoret. Probab., **29**:1 (2016), 32–47. MR3463076
[11] H.-K. Hwang, S. Janson, *Local limit theorems for finite and infinite urn models*, Ann. Probab., **36**:3 (2008), 992–1022. MR2408581
[12] S. Karlin, *Central limit theorems for certain infinite urn schemes*, J. Math. Mech., **17**:4 (1967), 373–401. MR0216548
[13] S.H. Sung, *An analogue of Kolmogorov's law of the iterated logarithm for arrays*, Bull. Austral. Math. Soc., **54**:2 (1996), 177–182. MR1411527
[14] N.S. Zakrevskaya, A.P. Kovalevskii, *One-parameter probabilistic models of text statistics*, Sib. Zh. Ind. Mat., **4**:2 (2001), 142–153. MR1965927

Mikhail Chebunin
Karlsruhe Institute of Technology,
Institute of Stochastics,
Karlsruhe, 76131, Germany
Novosibirsk State University,
2, Pirogova str.,
Novosibirsk, 630090, Russia.
Email address: chebuninmikhail@gmail.com