S(e)MR ISSN 1813-3304

СИБИРСКИЕ ЭЛЕКТРОННЫЕ
МАТЕМАТИЧЕСКИЕ ИЗВЕСТИЯ

Siberian Electronic Mathematical Reports

http://semr.math.nsc.ru

Том 20, №2, стр. 913–922 (2023) УДК 519.233

DOI 10.33048/semi.2023.20.055 MSC 62F03

LIMIT THEOREMS FOR FORWARD AND BACKWARD PROCESSES OF NUMBERS OF NON-EMPTY URNS IN INFINITE URN SCHEMES

M.G. CHEBUNIN, A.P. KOVALEVSKII

Abstract.

We study the joint asymptotics of forward and backward processes of numbers of non-empty urns in an infinite urn scheme. The probabilities of balls hitting the urns are assumed to satisfy the conditions of regular decrease. We prove weak convergence to a two-dimensional Gaussian process. Its covariance function depends only on exponent of regular decrease of probabilities. We obtain parameter estimates that have a normal asymototics for its joint distribution together with forward and backward processes. We use these estimates to construct statistical tests for the homogeneity of the urn scheme on the number of thrown balls.

Keywords: Zipf's law, weak convergence, Gaussian process, statistical test.

1. Introduction

Let $X_1, \ldots, X_n$ be independent and identically distributed positive integer-valued random variables,

$$p_i = \mathbf{P}(X_1 = i), \quad \sum_{i=1}^{\infty} p_i = 1. \tag{1}$$

Chebunin, M.G., Kovalevskii, A.P., Limit theorems for forward and backward processes of numbers of non-empty urns in infinite urn schemes.

The work is supported by Mathematical Center in Akademgorodok under agreement No. 075-15-2022-282 with the Ministry of Science and Higher Education of the Russian Federation.

Received November, 1, 2022, published November, 12, 2023.

The number of different elements among first k ones ($2 \le k \le n$) is

$$R_k = 1 + \sum_{i=2}^{k} \mathbf{1}(X_i \notin \{X_1, \dots, X_{i-1}\}). \tag{2}$$

Similarly, the number of different elements among *last* k ones ($2 \le k \le n$) is

$$R'_k = 1 + \sum_{i=n-k+1}^{n-1} \mathbf{1}(X_i \notin \{X_{n-k+2}, \dots, X_n\}). \tag{3}$$

In other words,

$$R'_k = 1 + \sum_{i=2}^{k} \mathbf{1}(X'_i \notin \{X'_1, \dots, X'_{i-1}\}), \tag{4}$$

with $X'_i = X_{n-i+1}$, the random variables in the backward order, $1 \le i \le n$.

We put by definition

$$R_0 = R'_0 = 0, \quad R_1 = R'_1 = 1. \tag{5}$$

Distributions of R_n and R'_n are identical, their limiting properties are known. We study their limiting joint disrtibution under the appropriate centering and normalizing.

If there is an infinite number of positive probabilities in (1) then this probability model is the infinite urn scheme. Karlin (1967) established the SLLN for R_n in the infinite urn scheme (Bahadur (1960) proved the weak LLN),

$$R_n / \mathbf{E}R_n \to 1 \quad \text{a.s.} \tag{6}$$

Now we need the regularity condition. Let $p_1 \ge p_2 \ge \ldots > 0$ and

$$\alpha(x) := \max\{k > 0 : \ p_k \ge 1/x\} = x^\theta L(x) \quad \text{as} \quad x \to \infty, \quad 0 < \theta < 1, \tag{7}$$

$L(\cdot)$ is the slowly varying function of the real argument: $L(tx)/L(x) \to 1$ as $x \to +\infty$ for any real $t > 0$.

From Karamata's characterization theorem, $\alpha(x)$ is a regulary varying function with index θ. The model (7) is the elementary probability model that corresponds to the Zipf's Law (Zipf, 1936) of power decreasing of word probabilities.

Karlin (1967) proved the CLT: if (7) holds then $(R_n - \mathbf{E}R_n)/\sqrt{\mathbf{Var}R_n}$ converges weakly to the standard normal distribution,

$$\mathbf{E}R_n \sim \Gamma(1-\theta)\alpha(n), \quad \mathbf{Var}R_n/\mathbf{E}R_n \to 2^\theta - 1, \tag{8}$$

$\Gamma(\cdot)$ is the Euler gamma.

From the Karlin's CLT and (8), $(R_n - \mathbf{E}R_n)/\sqrt{\mathbf{E}R_n}$ converges weakly to the centered normal distribution with variance $2^\theta - 1$. The CLT holds for $\theta = 1$ too but with another normalization.

Chebunin and Kovalevskii (2016) proved the Functional CLT: if (7) holds then the process

$$Z_n = \{Z_n(t),\ 0 \le t \le 1\} = \{(R_{[nt]} - \mathbf{E}R_{[nt]})/\sqrt{\mathbf{E}R_n},\ 0 \le t \le 1\} \tag{9}$$

converges weakly in $D(0,1)$ with uniform metric to a centered Gaussian process Z_θ with continuous a.s. sample paths and covariance function

$$K(s,t) = (s+t)^\theta - \max(s^\theta, t^\theta). \tag{10}$$

The Karlin's CLT is a particular case of the FCLT for $Z_n(1)$. The same FCLT is true for

$$Z'_n = \{Z'_n(t),\ 0 \le t \le 1\} = \{(R'_{[nt]} - \mathbf{E}R_{[nt]})/\sqrt{\mathbf{E}R_n},\ 0 \le t \le 1\}. \tag{11}$$

We prove the theorem about the joint limiting distribution of (Z_n, Z'_n).

All the papers on properties of R_n and similar statistics in the infinite urn scheme can be divided into 4 types:

1. Results under the regularity condition (7): the papers above, Durieu & Wang (2016), Chebunin (2017), Durieu, Samorodnitsky & Wang (2020), Chebunin & Zuyev (2020).
2. Results under (7) with $\theta = 0$ instead of $0 < \theta \le 1$, that is, for the slowly varying function $\alpha(x)$: Dutko (1989), Barbour (2009), Barbour & Gnedin (2009).
3. Results for the model without assuming the regularity condition (7): Key (1992, 1996), Hwang & Janson (2008), Muratov & Zuyev (2016), Ben-Hamou, Boucheron & Ohannessian (2017), Decrouez, Grabchak & Paris (2018).
4. Statistical applications — we postpone the survey of these results to Section 2.

Gnedin, Hansen & Pitman (2007) made a detalied survey of the results of types 1–3 existed at the time.

2. Main results

Theorem 1. *If (7) holds then the process* $(Z_n, Z'_n) = \{(Z_n(t), Z'_n(t)),\ 0 \le t \le 1\}$ *converges weakly in the uniform metric in* $D(0,1)^2$ *to 2-dimensional Gaussian process* (Z, Z') *with zero expectation and covariance function*

$$\mathbf{E}Z(s)Z(t) = \mathbf{E}Z'(s)Z'(t) = K(s,t), \quad \mathbf{E}Z(s)Z'(t) = K'(s,t),$$

where $K(s,t)$ *is given by (10), and*

$$K'(s,t) = ((s+t)^\theta - 1)\mathbf{1}(s+t > 1). \tag{12}$$

From Theorem 1 we have that the limiting process $\{(Z(t) - Z'(t))/\sqrt{2},\ 0 \le t \le 1/2\}$ is the stochastically self-similar process which coinside in distribution with the limiting process of Durieu and Wang (2016). So Theorem 1 gives an alternative way to simulate these processes without additional randomization.

We need some estimate of the unknown parameter θ to use the theorem in applications. Various classes of such estimates have been obtained and analysed by Hill (1975), Nicholls (1978), Zakrevskaya and Kovalevskii (2001, 2019), Guillou and Hall (2002), Ohannessian and Dahleh (2012), Chebunin (2014), Chebunin and Kovalevskii (2019a, 2019b), Chakrabarty et al. (2020).

But we need an estimate that is symmetric to the forward and backward processes. Moreover, we want to have the limiting joint distribution of the estimate and the two-dimensional process. We introduce the estimate and study its propetries in the next section.

3. Parameter's estimation

From (6) and (8), we have $\log R_n \sim \theta \log n$ a.s. Therefore, we may propose the following estimators for parameter θ:

$$\theta_n = \int_0^1 \log^+ R_{[nt]}\, dA(t), \quad \theta'_n = \int_0^1 \log^+ R'_{[nt]}\, dA(t),$$

here $\log^+ x = \max(\log x, 0)$. Function $A(\cdot)$ has bounded variation and

$$A(0) = A(1) = 0, \quad \lim_{x \downarrow 0} \log x \int_0^x |dA(t)| = 0, \quad \int_0^1 \log t \, dA(t) = 1. \tag{13}$$

Let

$$\widehat{\theta} = (\theta_n + \theta'_n)/2.$$

Theorem 2. *Let $p_i = i^{-1/\theta} l(i, \theta)$, $\theta \in [0,1]$, and $l(x, \theta)$ is a slowly varying function as $x \to \infty$. Then the estimator $\widehat{\theta}$ is strongly consistent.*

Proof. The proof follows from the definition of $\widehat{\theta}$ and Theorem 1 from Chebunin and Kovalevskii (2019). □

We need extra conditions to obtain the asymptotic normality of $\widehat{\theta}$.

Theorem 3. *Let $p_i = ci^{-1/\theta}(1 + o(i^{-1/2}))$, $\theta \in (0,1)$, and $A(t) = 0$, $t \in [0, \delta]$ for some $\delta \in (0, 1)$. Then*

$$\sqrt{\mathbf{E}R_n}(\widehat{\theta} - \theta) - \frac{1}{2} \int_0^1 t^{-\theta}(Z_n(t) + Z'_n(t)) \, dA(t) \to_p 0.$$

Proof. The proof follows from the definition of $\widehat{\theta}$ and Theorem 2 from Chebunin and Kovalevskii (2019). □

From Theorem 3, it follows that $\widehat{\theta}$ converges to θ at rate $(\mathbf{E}R_n)^{-1/2}$, and $\sqrt{\mathbf{E}R_n}(\widehat{\theta} - \theta)$ converges weakly to the normal random variable $\frac{1}{2} \int_0^1 t^{-\theta}(Z_\theta(t) + Z'_\theta(t)) \, dA(t)$ with variance $\frac{1}{2} \int_0^1 \int_0^1 (st)^{-\theta}(K(s,t) + k(s,t)) \, dA(s) \, dA(t)$.

Example 1 Take

$$A(t) = \begin{cases} 0, & 0 \le t \le 1/2; \\ -(\log 2)^{-1}, & 1/2 < t < 1; \\ 0, & t = 1. \end{cases}$$

Then

$$\widehat{\theta} = \log_2 \left(R_n / \sqrt{R_{[n/2]} R'_{[n/2]}} \right), \quad n \ge 2.$$

4. Test for a known rate

Let $0 < \theta < 1$ be known. We introduce *empirical bridges* $\overset{o}{Z}_n$, $\overset{o}{Z'}_n$ (Kovalevskii and Shatalin, 2015, 2016) as follows.

$$\overset{o}{Z}_n(k/n) = \left(R_k - (k/n)^\theta R_n\right) / \sqrt{R_n}, \quad \overset{o}{Z'}_n(k/n) = \left(R'_k - (k/n)^\theta R_n\right) / \sqrt{R_n},$$

$0 \le k \le n$, where $R_0 = 0$. We construct a piecewise linear approximation: for any $0 \le u < 1/n$ and $0 \le k \le n - 1$,

$$\overset{o}{Z}_n\left(\frac{k}{n} + u\right) = \overset{o}{Z}_n(k/n) + nu\left(\overset{o}{Z}_n((k+1)/n) - \overset{o}{Z}_n(k/n)\right),$$

$$\overset{o}{Z'}_n\left(\frac{k}{n} + u\right) = \overset{o}{Z'}_n(k/n) + nu\left(\overset{o}{Z'}_n((k+1)/n) - \overset{o}{Z'}_n(k/n)\right).$$

Theorem 4. *Under the assumptions of Theorem 2,*

$$\sup_{0\le t\le 1} |\overset{o}{Z}_n(t) - (Z_n(t) - t^\theta Z_n(1))| \to 0 \ a.s.$$

$$\sup_{0\le t\le 1} |\overset{o}{Z'}_n(t) - (Z'_n(t) - t^\theta Z'_n(1))| \to 0 \ a.s.$$

Proof. The first statement is Theorem 3 from Chebunin and Kovalevskii (2019). For the second statement, let $t \in [0,1)$, and $k = [nt]$, then $t = k/n + u$, $0 \le k \le n-1$, $u \in [0, 1/n)$. Let $f_\theta(x) = (1+x)^\theta - x^\theta$. So $0 \le f_\theta(x) \le f_\theta(0) = 1$ for $x \ge 0$.

By the definition of $\overset{o}{Z'}_n(t)$,

$$\frac{R'_k - \left(\frac{k+1}{n}\right)^\theta R_n}{\sqrt{R_n}} \le \overset{o}{Z'}_n(t) \le \frac{R'_{k+1} - \left(\frac{k}{n}\right)^\theta R_n}{\sqrt{R_n}},$$

so

$$\left| \overset{o}{Z'}_n(t) - \frac{R_{[nt]} - t^\theta R_n}{\sqrt{R_n}} \right| \le \frac{R'_{k+1} - R'_k + \frac{1}{n^\theta} f_\theta(k) R_n}{\sqrt{R_n}}$$

$$\le \frac{1}{\sqrt{R_n}} + \frac{\sqrt{R_n}}{n^\theta} \to 0$$

a.s. unformly on $t \in [0,1]$.

□

Let C(0,1) be the set of all continious functions on $[0,1]$ with the uniform metric $\rho(x,y) = \max_{t\in[0,1]} |x(t) - y(t)|$. By Theorem 1, we have

Corollary 1. *Under the assumptions of Theorem 4,* $(\overset{o}{Z}_n, \overset{o}{Z'}_n)$ *converges weakly in* $C(0,1)$ *to 2-dimensional Gaussian process* $(\overset{o}{Z}, \overset{o}{Z'})$ *that can be represented as* $(\overset{o}{Z}(t), \overset{o}{Z'}(t)) = (Z_\theta(t) - t^\theta Z_\theta(1), Z'_\theta(t) - t^\theta Z'_\theta(1))$, $0 \le t \le 1$. *Its correlation function is given by the covariance function*

$$c_{R,R}(s,t) = c_{R',R'}(s,t) = \overset{o}{K}(s,t), \quad c_{R,R'}(s,t) = \overset{o}{K'}(s,t),$$

where

$$\overset{o}{K}(s,t) = K(s,t) - s^\theta K(1,t) - t^\theta K(s,1) + s^\theta t^\theta K(1,1),$$

$$\overset{o}{K'}(s,t) = K'(s,t) - s^\theta K'(1,t) - t^\theta K'(s,1) + s^\theta t^\theta K'(1,1).$$

Now we show how to implement the goodness-of-fit test in this case.

Let $W_n^2 = \int\limits_0^1 \left(\overset{o}{Z}_n(t)\right)^2 + \left(\overset{o}{Z'}_n(t)\right)^2 dt$. It is equal to

$$W_n^2 = \frac{1}{3n} \sum_{k=1}^{n-1} \overset{o}{Z}_n\left(\frac{k}{n}\right)\left(2\overset{o}{Z}_n\left(\frac{k}{n}\right) + \overset{o}{Z}_n\left(\frac{k+1}{n}\right)\right) \tag{14}$$

$$+\frac{1}{3n} \sum_{k=1}^{n-1} \overset{o}{Z'}_n\left(\frac{k}{n}\right)\left(2\overset{o}{Z'}_n\left(\frac{k}{n}\right) + \overset{o}{Z'}_n\left(\frac{k+1}{n}\right)\right).$$

Then W_n^2 converges weakly to $W_\theta^2 = \int\limits_0^1 \left(\overset{o}{Z}_\theta(t)\right)^2 + \left(\overset{o}{Z'}_\theta(t)\right)^2 dt$.

So the test rejects the null hypothesis if $W_n^2 \geq C$. The p-value of the test is $1 - F_\theta(W_{n,obs}^2)$. Here F_θ is the cumulative distribution function of W_θ^2 and $W_{n,obs}^2$ is the value of W_n^2 for observations under consideration.

One can estimate F_θ by simulations or find it explicitely using the Smirnov formula (Smirnov, 1937): if $W_\theta^2 = \sum_{k=1}^\infty \frac{\eta_k^2}{\lambda_k}$, $\eta_1, \eta_2, \ldots$ are independent and have standard normal distribution, $0 < \lambda_1 < \lambda_2 < \ldots$, then

$$F_\theta(x) = 1 + \frac{1}{\pi}\sum_{k=1}^\infty (-1)^k \int_{\lambda_{2k-1}}^{\lambda_{2k}} \frac{e^{-\lambda x/2}}{\sqrt{-D(\lambda)}} \cdot \frac{d\lambda}{\lambda},\ x > 0, \tag{15}$$

$$D(\lambda) = \prod_{k=1}^\infty \left(1 - \frac{\lambda}{\lambda_k}\right).$$

The integrals in the RHS of (15) must tend to 0 monotonically as $k \to \infty$, and λ_k^{-1} are the eigenvalues of the kernel (see Martynov (1973), Chapter 3).

5. Test for an unknown rate

Let us introduce the process $(\widehat{Z}_n, \widehat{Z'}_n)$:

$$\widehat{Z}_n(k/n) = \left(R_k - (k/n)^{\widehat{\theta}} R_n\right)/\sqrt{R_n},\quad \widehat{Z'}_n(k/n) = \left(R'_k - (k/n)^{\widehat{\theta}} R_n\right)/\sqrt{R_n},$$

$0 \leq k \leq n$. As for $\overset{o}{Z}_n$, let for $0 \leq u < 1/n$ and $0 \leq k \leq n-1$

$$\widehat{Z}_n\left(\frac{k}{n} + u\right) = \widehat{Z}_n(k/n) + nu\left(\widehat{Z}_n((k+1)/n) - \widehat{Z}_n(k/n)\right),$$

$$\widehat{Z'}_n\left(\frac{k}{n} + u\right) = \widehat{Z'}_n(k/n) + nu\left(\widehat{Z'}_n((k+1)/n) - \widehat{Z'}_n(k/n)\right).$$

Theorem 5. *Under assumptions of Theorem 3, $(\widehat{Z}_n, \widehat{Z'}_n)$ converges weakly as $n \to \infty$ to 2-dimensional Gaussian process $(\widehat{Z}_\theta, \widehat{Z'}_\theta)$ that can be represented as $(\widehat{Z}_\theta(t), \widehat{Z'}_\theta(t))$, $0 \leq t \leq 1$, where*

$$\widehat{Z}_\theta(t) = \overset{o}{Z}(t) - \frac{t^\theta \log t}{2}\int_0^1 u^{-\theta}(Z_\theta(u) + Z'_\theta(u))\,dA(u),$$

$$\widehat{Z'}_\theta(t) = \overset{o}{Z'}(t) - \frac{t^\theta \log t}{2}\int_0^1 u^{-\theta}(Z_\theta(u) + Z'_\theta(u))\,dA(u).$$

Proof. Similarly to the proof of Theorem 4 from Chebunin and Kovalevskii (2019), we can show that

$$\sup_{t\in[0,1]}\left|\widehat{Z}_n(t) - (\overset{o}{Z}_n(t) - \sqrt{R_n}(\widehat{\theta} - \theta)t^\theta \log t)\right| \to_p 0,$$

$$\sup_{t\in[0,1]}\left|\widehat{Z'}_n(t) - (\overset{o}{Z'}_n(t) - \sqrt{R_n}(\widehat{\theta} - \theta)t^\theta \log t)\right| \to_p 0.$$

Let us demonstrate it for $\widehat{Z'}_n(t)$. Let $t \in [0,1)$, $k = [nt]$, $u = t - k/n$, $f_\theta(x) = (1+x)^\theta - x^\theta$ as in the proof of Theorem 4. By the definition,

$$\widehat{Z'}_n(k/n) = \overset{o}{Z'}_n(k/n) + \sqrt{R_n}\left((k/n)^\theta - (k/n)^{\widehat{\theta}}\right),$$

$$\widehat{Z'}_n(t) = \overset{o}{Z'}_n(t) + \sqrt{R_n}\left((k/n)^\theta - (k/n)^{\widehat{\theta}}\right)$$

$$+nu\sqrt{R_n}\left(\left(\frac{k+1}{n}\right)^{\theta}-\left(\frac{k+1}{n}\right)^{\widehat{\theta}}-\left(\frac{k}{n}\right)^{\theta}+\left(\frac{k}{n}\right)^{\widehat{\theta}}\right).$$

We have

$$\left(\frac{k+1}{n}\right)^{\theta}-\left(\frac{k}{n}\right)^{\theta}=f_\theta(k)/n^\theta,\quad \left(\frac{k+1}{n}\right)^{\widehat{\theta}}-\left(\frac{k}{n}\right)^{\widehat{\theta}}=f_{\widehat{\theta}}(k)/n^{\widehat{\theta}},$$

so

$$\left|\widehat{Z'}_n(t)-\overset{o}{Z'}_n(t)+\sqrt{R_n}\left(t^{\widehat{\theta}}-t^\theta\right)\right|$$
$$=\left|\widehat{Z'}_n(t)-\overset{o}{Z'}_n(t)+\sqrt{R_n}\left(\left(\frac{k}{n}+u\right)^{\widehat{\theta}}-\left(\frac{k}{n}+u\right)^{\theta}\right)\right|$$
$$\le 2\sqrt{R_n}\left(f_\theta(k)/n^\theta+f_{\widehat{\theta}}(k)/n^{\widehat{\theta}}\right)\le 2\sqrt{R_n}\left(1/n^\theta+1/n^{\widehat{\theta}}\right)\to 0$$

a.s. unformly in $t\in[0,1]$.

Note that one can change $t^{\widehat{\theta}}-t^\theta$ by $(\widehat{\theta}-\theta)t^\theta\log t$. Indeed,

$$t^{\widehat{\theta}}-t^\theta=t^\theta\left(e^{(\widehat{\theta}-\theta)\log t}-1\right)$$
$$=(\widehat{\theta}-\theta)t^\theta\log t+t^\theta\sum_{k\ge 2}\frac{((\widehat{\theta}-\theta)\log t)^k}{k!}$$
$$=(\widehat{\theta}-\theta)t^\theta\log t+t^\theta(\widehat{\theta}-\theta)^2(1+o(1))\sum_{k\ge 2}\frac{\log^k t}{k!}$$
$$=(\widehat{\theta}-\theta)t^\theta\log t\left(1+(\widehat{\theta}-\theta)(1+o(1))\frac{e^{\log t}-1-\log t}{\log t}\right)$$
$$=(\widehat{\theta}-\theta)t^\theta\log t(1+o(1))$$

a.s. unformly in $t\in[0,1]$. Hence from Theorems 3 and 4, we have a joint weak convergence of

$$(\overset{o}{Z}_n,\overset{o}{Z'}_n,\sqrt{R_n}(\widehat{\theta}-\theta))$$

to

$$(\overset{o}{Z}_\theta,\overset{o}{Z'}_\theta,\ \frac{1}{2}\int_0^1 u^{-\theta}(Z_\theta(u)+Z'_\theta(u))\,dA(u)).$$

So, $(\widehat{Z}_n,\widehat{Z'}_n)$ converges weakly to $(\widehat{Z}_\theta,\widehat{Z'}_\theta)$.

□

Corollary 2. *Assume the conditions of Theorem 2 to hold. Let* $\widehat{W}_n^2=\int\limits_0^1\left(\widehat{Z}_n(t)\right)^2+\left(\widehat{Z'}_n(t)\right)^2dt$. *Then* $\widehat{W}_n^2$ *converges weakly to* $\widehat{W}_\theta^2=\int\limits_0^1\left(\widehat{Z}_\theta(t)\right)^2+\left(\widehat{Z'}_\theta(t)\right)^2dt$.

Similarly to (14), $\widehat{W}_n^2$ has the following representation

$$\widehat{W}_n^2=\frac{1}{3n}\sum_{k=1}^{n-1}\widehat{Z}_n\left(\frac{k}{n}\right)\left(2\widehat{Z}_n\left(\frac{k}{n}\right)+\widehat{Z}_n\left(\frac{k+1}{n}\right)\right)$$
$$+\frac{1}{3n}\sum_{k=1}^{n-1}\widehat{Z'}_n\left(\frac{k}{n}\right)\left(2\widehat{Z'}_n\left(\frac{k}{n}\right)+\widehat{Z'}_n\left(\frac{k+1}{n}\right)\right)$$

The p-value of the goodness-of fit test is $1-\widehat{F}_\theta(\widehat{W}^2_{n,obs})$. Here $\widehat{F}_\theta$ is the cumulative distribution function of $\widehat{W}^2_\theta$, and $\widehat{W}^2_{n,obs}$ is the observed value of $\widehat{W}^2_n$. Further, the function $\widehat{F}_\theta$ can be found using the approach described in Section 3 with λ_k replaced by the eigenvalues $\widehat{\lambda}_k$ of the kernel in the Smirnov formula.

6. Proof of Theorem 1

We denote by $\mathbb{X}_i(n)$ the number of balls in urn i. Let $\Pi = \{\Pi(t), t \geq 0\}$ be a Poisson process with parameter 1. The Poissonized version of Karlin model assumes the total number $\Pi(n)$ of balls. According to the well-known thinning property of Poisson flows, stochastic processes $\left\{\mathbb{X}_i(\Pi(t)) \stackrel{\text{def}}{=} \mathbf{\Pi}_i(t), t \geq 0\right\}$ are Poisson processes (with intensities p_i) which are mutually independent for different i 's. The definition implies that for any fixed $n \geq 1$, $\tau, t \in [0,1]$

$$R_{\Pi(tn)} = \sum_{k=1}^{\infty} \mathbf{I}\left(\mathbf{\Pi}_k(tn) > 0\right) = \sum_{k=1}^{\infty} \mathbf{I}_k(tn),$$

$$R'_{\Pi(\tau n)} = \sum_{k=1}^{\infty} \mathbf{I}\left(\mathbf{\Pi}_k(n) - \mathbf{\Pi}_k((1-\tau)n) > 0\right) = \sum_{k=1}^{\infty} \mathbf{I}'_k(\tau n).$$

Step 1 (covariances) Let $\tau, t \in [0,1]$

$$\mathbf{cov}\left(R_{\Pi(tn)}, R'_{\Pi(\tau n)}\right) = \sum_{k=1}^{\infty} \mathbf{cov}(\mathbf{I}_k(tn), \mathbf{I}'_k(\tau n))$$

$$= \sum_{k=1}^{\infty} \left(\mathbf{P}\left(\mathbf{\Pi}_k(tn) > 0, \mathbf{\Pi}_k(n) - \mathbf{\Pi}_k((1-\tau)n) > 0\right) - (1-e^{-p_k tn})(1-e^{-p_k \tau n})\right)$$

Note that if $t+\tau > 1$, then

$$\begin{aligned}
\mathbf{P}(\mathbf{\Pi}_k(tn) > 0, \mathbf{\Pi}_k(n) - \mathbf{\Pi}_k((1-\tau)n) > 0) &= \mathbf{P}(\mathbf{\Pi}_k(tn) - \mathbf{\Pi}_k((1-\tau)n) > 0) \\
+\mathbf{P}(\mathbf{\Pi}_k(tn) - \mathbf{\Pi}_k((1-\tau)n) = 0, \mathbf{\Pi}_k((1-\tau)n) > 0, \mathbf{\Pi}_k(n) - \mathbf{\Pi}_k(tn) > 0) \\
= 1 - e^{-p_k(t+\tau-1)n} + e^{-p_k(t+\tau-1)n}(1-e^{-p_k(1-\tau)n})(1-e^{-p_k(1-t)n}) \\
= 1 - e^{-p_k tn} - e^{-p_k \tau n} + e^{-p_k n}.
\end{aligned}$$

Hence

$$\mathbf{cov}\left(R_{\Pi(tn)}, R'_{\Pi(\tau n)}\right) = \mathbf{I}(t+\tau > 1) \sum_{k=1}^{\infty} \left(e^{-p_k(t+\tau)n} - e^{-p_k n}\right)$$

$$= \mathbf{I}(t+\tau > 1)(\mathbf{E}R_{\Pi((t+\tau)n} - \mathbf{E}R_{\Pi(n)}).$$

Since

$$\mathbf{E}R_{\Pi(tn)}/\mathbf{E}R_n \sim t^\theta,$$

then

$$\mathbf{cov}\left(R_{\Pi(tn)}, R'_{\Pi(\tau n)}\right)/\mathbf{E}R_n \sim K'(t,\tau).$$

Step 2 (convergence of finite-dimensional distributions) Analogously to the proof of Theorem 1 in Dutko (1989), we have that, for any fixed $m \geq 1, 0 < t_1 < t_2 < \cdots < t_m \leq 1$, the triangle array of $2m$ -dimensional random vectors $\left\{((\mathbf{I}_k(nt_i) - \mathbf{E}\mathbf{I}_k(nt_i))/\sqrt{\mathbf{E}R_n}, (\mathbf{I}'_k(nt_i) - \mathbf{E}\mathbf{I}'_k(nt_i))/\sqrt{\mathbf{E}R_n}, i \leq m), k \leq n\right\}_{n\geq 1}$ satisfies the Lindeberg condition (see Borovkov (2013), Theorem 8.6.2, p. 215).

Step 3 (relative compactness) Since $R_{\Pi(nt)} \overset{d}{=} R'_{\Pi(nt)}$, then the relative compactness follows from Chebunin and Kovalevskii (2016).

Step 4 (approximation of the original process)

It follows from the corresponding step of the proof in Chebunin and Kovalevskii (2016) and the previous step.

Theorem 1 is proved.

Acknowledgements

The work is supported by Mathematical Center in Akademgorodok under agreement No. 075-15-2022-282 with the Ministry of Science and Higher Education of the Russian Federation. The authors thank the anonymous referee for very helpful comments that improved the quality of the manuscript.

References

[1] R.R. Bahadur, *On the number of distinct values in a large sample from an infinite discrete distribution*, Proc. Natl. Inst. Sci. India, Part A, Supp. II, **26** (1960), 67–75. Zbl 0151.23803

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

[3] A.D. Barbour, A.V. Gnedin, *Small counts in the infinite occupancy scheme*, Electron. J. Probab., **14** (2009), Paper no. 13, 365–384. MR2480545

[4] 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. Zbl 1366.60016

[5] A. Chakrabarty, M. Chebunin, A. Kovalevskii, I. Pupyshev, N. Zakrevskaya, Q. Zhou, *A statistical test for correspondence of texts to the Zipf-Mandelbrot law*, Sib. Èlektron. Mat. Izv., **17** (2020), 1959–1974. Zbl 1459.62028

[6] M.G. Chebunin, *Estimation of parameters of probabilistic models which is based on the number of different elements in a sample*, Sib. Zh. Ind. Mat., **17**:3 (2014), 135–147. Zbl 1340.62018

[7] M.G. Chebunin, *Functional central limit theorem in an infinite urn scheme for distributions with superheavy tails*, Sib. Èlektron. Mat. Izv., **14** (2017), 1289–1298. Zbl 1391.60058

[8] M. Chebunin, A. Kovalevskii, *Functional central limit theorems for certain statistics in an infinite urn scheme*, Stat. Probab. Lett., **119** (2016), 344–348. Zbl 1398.60051

[9] M. Chebunin, A. Kovalevskii, *A statistical test for the Zipf's law by deviations from the Heaps' law*, Sib. Èlektron. Mat. Izv., **16** (2019), 1822–1832. Zbl 1433.62060

[10] M. Chebunin, A. Kovalevskii, *Asymptotically normal estimators for Zipf's law*, Sankhyā, Ser. A, **81**:2 (2019), 482–492. Zbl 1437.62097

[11] M. Chebunin, S. Zuyev, *Functional central limit theorems for occupancies and missing mass process in infinite urn models*, J. Theor. Probab., **35**:1 (2020), 1–19. Zbl 1484.60039

[12] G. Decrouez, M. Grabchak, Q. Paris, *Finite sample properties of the mean occupancy counts and probabilities*, Bernoulli, **24**:3 (2018), 1910–1941. Zbl 1429.60016

[13] O. Durieu, Y. Wang, *From infinite urn schemes to decompositions of self-similar Gaussian processes*, Electron. J. Probab., **21** (2016), Paper No. 43. Zbl 1346.60039

[14] O. Durieu, G. Samorodnitsky, Y. Wang, *From infinite urn schemes to self-similar stable processes*, Stochastic Processes Appl., **130**:4 (2020), 2471–2487. Zbl 1434.60105

[15] M. Dutko, *Central limit theorems for infinite urn models*, Ann. Probab. **17**:3 (1989), 1255–1263. Zbl 0685.60023

[16] A. Guillou, P. Hall, *A diagnostic for selecting the threshold in extreme value analysis*, J. R. Stat. Soc., Ser. B, **63**:2 (2001), 293–305. Zbl 0979.62039

[17] 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. Zbl 1189.60050

[18] B.M. Hill, *A simple general approach to inference about the tail of a distribution*, Ann. Stat., **3**:5 (1975), 1163–1174. Zbl 0323.62033

[19] H.-K. Hwang, S. Janson, *Local limit theorems for finite and infinite urn models*, Ann. Probab., **36**:3 (2008), 992–1022. Zbl 1138.60027
[20] S. Karlin, *Central limit theorems for certain infinite urn schemes*, J. Math. Mech., **17**:4 (1967), 373–401. Zbl 0154.43701
[21] E.S. Key, *Rare numbers*, J. Theor. Probab., **5**:2 (1992), 375–389. Zbl 0763.60013
[22] E.S. Key, *Divergence rates for the number of rare numbers*, J. Theor. Probab., **9**:2 (1996), 413–428. Zbl 0870.60023
[23] A. Muratov, S. Zuyev, *Bit flipping and time to recover*, J. Appl. Probab., **53**:3 (2016), 650–666. Zbl 1351.60104
[24] P.T. Nicholls, *Estimation of Zipf parameters*, J. Am. Soc. Inf. Sci., **38**:6 (1987), 443–445.
[25] M.I. Ohannessian, M.A. Dahleh, *Rare probability estimation under regularly varying heavy tails*, Proceedings of the 25th Annual Conference on Learning Theory, PMLR, **23**, 2012, 21.1–21.24.
[26] N.S. Zakrevskaya, A.P. Kovalevskii, *One-parameter probabilistic models for text statistics*, Sib. Zh. Ind. Mat., **4**:2 (2001), 142–153. Zbl 1031.68123
[27] N. Zakrevskaya, A. Kovalevskii, *An omega-square statistics for analysis of correspondence of small texts to the Zipf-Mandelbrot law*, in *Applied methods of statistical analysis. Statistical computation and simulation — AMSA'2019*, Proceedings of the International Workshop, NSTU, Novosibirsk, 2019, 488–494.
[28] G.K. Zipf, *The psycho-biology of language*, Routledge, London, 1936.

Mikhail Georgievich Chebunin
Karlsruhe Institute of Technology,
Institute of Stochastics,
76131, Karlsruhe, Germany;
Novosibirsk State University,
Pirogova str., 1,
630090, Novosibirsk, Russia
Email address: chebuninmikhail@gmail.com

Artyom Pavlovich Kovalevskii
Sobolev Institute of Mathematics,
Koptyuga pr., 4,
Novosibirsk State University,
Pirogova str., 1,
630090, Novosibirsk, Russia;
Novosibirsk State Technical University,
K. Marksa ave., 20,
630073, Novosibirsk, Russia
Email address: artyom.kovalevskii@gmail.com