Sankhyā : The Indian Journal of Statistics
2019, Volume 81-A, Part 2, pp. 482-492

Asymptotically Normal Estimators for Zipf's Law

Mikhail Chebunin
Sobolev Institute of Mathematics, Novosibirsk, Russia
Novosibirsk State University, Novosibirsk, Russia
Artyom Kovalevskii
Novosibirsk State Technical University, Novosibirsk, Russia
Novosibirsk State University, Novosibirsk, Russia

Abstract

We study an infinite urn scheme with probabilities corresponding to a power function. Urns here represent words from an infinitely large vocabulary. We propose asymptotically normal estimators of the exponent of the power function. The estimators use the number of different elements and a few similar statistics. If we use only one of the statistics we need to know asymptotics of a normalizing constant (a function of a parameter). All the estimators are implicit in this case. If we use two statistics then the estimators are explicit, but their rates of convergence are lower than those for estimators with the known normalizing constant.

AMS (2000) *subject classification.* Primary 62F10; Secondary 62F12.
Keywords and phrases. Infinite urn scheme, Zipf's law, Asymptotic normality.

1 Introduction

Zipf's law (Zipf, 1949) states that sequential frequencies f_i of words in a text equal $ci^{-1/\theta}$, $c > 0$, $\theta \in (0, 1)$, $i > i_0 \geq 0$. Its modification is Mandelbrot's law (Mandelbrot, 1965) that states that $f_i = c(i + \beta)^{-1/\theta}$, $\beta \geq 0$.

Probabilistic interpretation of these and similar laws is an infinite urn scheme studied by Bahadur (1960) & Karlin (1967). There are n balls that are distributed to urns independently and randomly; there are infinitely many urns. Each ball goes to urn i with probability $p_i > 0$, $p_1 + p_2 + \ldots = 1$ (frequencies converge a.s. to probabilities).

So, urns here represent words from an infinitely large vocabulary and balls represent consecutive words of a text. In this model words in a text are independent and match i-th word in the vocabulary with probability p_i.

We assume that $p_1 \geq p_2 \geq \ldots$ and one of the following asymptotics holds (the second one is wider than the first):

$$p_i = c i^{-1/\theta}(1 + o(i^{-1/2})), \tag{1.1}$$

$\theta \in (0,1)$, $c = c(\theta)$ (this assumption includes Zipf's and Mandelbrot's laws);

$$p_i = i^{-1/\theta} L_0(i, \theta), \tag{1.2}$$

$L_0(x, \theta)$ is a slowly varying function of x in Karamata's sense for any fixed $\theta \in (0,1)$.

Our aim is to construct asymptotically normal estimators of θ under (1.1). We state its strong consistency under (1.2). To do so we use statistics studied by Bahadur (1960), Karlin (1967), Dutko (1989), Key (1992, 1996), Zakrevskaya and Kovalevskii (2001), Gnedin et al. (2007), Boonta and Neammanee (2007), Hwang and Janson (2008), Bogachev et al. (2008), Barbour (2009), Barbour and Gnedin (2009), Ohannessian and Dahleh (2012), Chebunin (2014), Chebunin and Kovalevskii (2016), Muratov and Zuyev (2016), Ben-Hamou et al. (2017).

Nicholls (1987) collected a few classes of estimators and tested them on sciencemetric data. But asymptotical normality of any of estimators had not been proved. But one needs an asymptotic normality to calculate inference for hypothesis of homogeneity of two texts. Our theorems state the neccessary convergencies and therefore give approaches to testing the homogeneity of texts.

Let $J_i(n)$ be the number of balls in the ith urn, R_n be the number of non-empty urns, and $R^*_{n,k}$ be the number of urns with not less than $k \geq 1$ balls

$$R_n = \sum_{i=1}^{\infty} \mathbf{I}\{J_i(n) > 0\}, \quad R^*_{n,k} = \sum_{i=1}^{\infty} \mathbf{I}(J_i(n) \geq k).$$

Note that $R^*_{n,1} = R_n$. The number of urns with exactly k balls: $R_{n,k} = R^*_{n,k} - R^*_{n,k+1}$. The number of urns with odd number of balls:

$$U_n = \sum_{i=1}^{\infty} \mathbf{I}(J_i(n) \equiv 1 (\mathrm{mod}\ 2)).$$

Karlin (1967) suggested studying a random sample with a random number of experiments $\Pi(t)$. Here $\{\Pi(t),\ t \geq 0\}$ is a Poisson process with parameter 1. Random choice of an urn and Poisson process are independent. Processes $\{J_i(\Pi(t)) \stackrel{def}{=} \Pi_i(t),\ t \geq 0\}$ are independent Poisson processes with parameters p_i. Apart from being in the listed papers, the Poissonization is used by Ben-Hamou et al. (2016) for estimating codes on countable alphabets, by Durieu and Wang (2016) in proof of functional CLT for some randomization of statistics R_n and U_n, by Grubel and Hitczenko (2009) in studying limit distributions of gaps in discrete random samples, by Khmaladze (2011) for more general allocation schemes.

From definition,

$$R^*_{\Pi(t),k} = \sum_{i=1}^{\infty} \mathbf{I}(\Pi_i(t) \geq k), \;\; R_{\Pi(t),k} = \sum_{i=1}^{\infty} \mathbf{I}(\Pi_i(t) = k),$$

$$U_{\Pi(t)} = \sum_{i=1}^{\infty} \mathbf{I}(\Pi_i(t) \equiv 1 (\mathrm{mod}\ 2)).$$

Karlin (1967) introduced function $\alpha(x) = \max\{j|\ p_j \geq 1/x\}$ and proved that (1.2) implies $\alpha(x) = x^{\theta} L(x, \theta)$, and $L(x, \theta)$ is a slowly varying function as $x \to \infty$.

Karlin proved SLLNs for all the statistics under (1.2). Karlin proved CLTs for R_n, U_n and vector $(R_{n,1}, \dots, R_{n,d})$ for any finite d.

Karlin proved that asymptotics of expectations of all of the statistics are proportional to $\alpha(n)$ with some coefficient depending on θ only. This law was found for texts empirically (with $L(x, \theta) = L(\theta)$) by Herdan (1960) and Heaps (1978, Sect. 3.7). It is interesting that modern large-scale studies of languages show a deviation from this law (Petersen et al., 2012) that is interpreted as a decrease in need of acquiring new words.

The authors do not know of any estimator of θ with proved asymptotic normality. An estimator by Zakrevskaya and Kovalevskii (2001) found by a substitution method is (we will see it) asymptotically normal for Zipf's law but authors proved consistency only. An estimator of Chebunin (2014) is strongly consistent but is not asymptotically normal. We will prove asymptotic normality of estimators of Ohannessian and Dahleh (2012) under (1.1) but authors proved only strong consistency under (1.2).

The rest of the paper is organized as follows. In Section 2 we construct asymptotically normal estimators of θ using only one of the statistics. This is possible only if constant C is known (it can be a differentiable function of θ) in (1.1), and all the estimators are implicit in this case. In Section 3 we prove asymptotic normality of estimators based on two statistics. We use multidimensional CLTs for $(R_{n,1}, \dots, R_{n,d})$ proved by Karlin (1967) and for $(R_n, R_{n,1}, \dots, R_{n,d})$ proved in Appendix in a functional generalization.

We use designation $\Rightarrow \mathbf{N}_{0,\sigma^2}$ for weak convergence to a normal distribution with zero mean and variance σ^2. All convergencies are under $n \to \infty$.

2 Implicit Estimators Using One Statistics

We prove a general theorem for some abstract statistics S_n in infinite urn scheme with required properties. Then we prove that these properties are satisfied for all statistics under consideration if one assumes (1.1).

Let $S_n/n^\theta l(n,\theta) \overset{a.s.}{\to} 1$ as $n\to\infty$, where $l(\theta,n)$ is a slowly varying function. Let us define $\theta_n^*\in(0,1)$ as a solution of the equation

$$S_n = n^\theta l(\theta,n). \tag{2.1}$$

As $\ln S_n - \theta\ln n - \ln l(\theta,n)\to 0$, so

$$\frac{\ln S_n}{\ln n}\overset{a.s.}{\to}\theta,\quad \text{and}\quad \frac{\ln S_n}{\ln n}-\theta_n^* = \frac{\ln l(\theta_n^*,n)}{\ln n}\overset{a.s.}{\to} 0.$$

So θ_n^* is a strongly consistent estimator of θ. We will study asymptotic normality of θ_n^*. Let

$$\mathbf{E}S_n = n^\theta l(\theta,n)+o(\sqrt{\mathbf{E}S_n}),\ \ \frac{\mathbf{Var}S_n}{\mathbf{E}S_n}\to\sigma^2,\ \ \frac{S_n}{\mathbf{E}S_n}\overset{a.s.}{\to}1,\ \ \frac{S_n-\mathbf{E}S_n}{\sqrt{\mathbf{Var}S_n}}\Rightarrow \mathbf{N}_{0,1}, \tag{2.2}$$

$l(\theta,n)$ is a slowly varying function as $n\to\infty$.

THEOREM 1. *Suppose* (2.2) *holds and*

$$\frac{\ln l(\theta_n^*,n)-\ln l(\theta,n)}{(\theta_n^*-\theta)\ln n}\overset{def}{=}\widetilde{l}_n\overset{p}{\to}0,$$

θ_n^* *is a solution of* (2.1). *Then*

$$\ln n\sqrt{S_n}(\theta_n^*-\theta)\Rightarrow \boldsymbol{N}_{0,\sigma^2}.$$

PROOF. $S_n^0 := \frac{S_n-n^\theta l(\theta,n)}{\sqrt{\mathbf{Var}S_n}}\Rightarrow\mathbf{N}_{0,1}$. From (2.2)

$$\ln S_n-\ln(n^\theta l(\theta,n)) = \ln\left(1+\frac{S_n}{n^\theta l(\theta,n)}-1\right)\overset{a.s.}{\sim}\frac{S_n}{n^\theta l(\theta,n)}-1$$

as $n\to\infty$. Then

$$S_n^0 = \frac{n^\theta l(\theta,n)}{\sqrt{\mathbf{Var}S_n}}\left(\frac{S_n}{n^\theta l(\theta,n)}-1\right)\overset{a.s.}{\sim}\sqrt{\frac{n^\theta l(\theta,n)}{\sigma^2}}\left(\frac{S_n}{n^\theta l(\theta,n)}-1\right)$$

$$\overset{a.s.}{\sim} \sqrt{\frac{S_n}{\sigma^2}}(\ln S_n - \theta \ln n - \ln l(\theta, n)) = \sqrt{\frac{S_n}{\sigma^2}}(\theta_n^* \ln n + \ln l(\theta_n^*, n) - \theta \ln n - \ln l(\theta, n))$$

$$= \ln n \sqrt{\frac{S_n}{\sigma^2}}(\theta_n^* - \theta)\left(1 + \frac{\ln l(\theta_n^*, n) - \ln l(\theta, n)}{(\theta_n^* - \theta)\ln n}\right)$$

$$\sim \ln n \sqrt{\frac{S_n}{\sigma^2}}(\theta_n^* - \theta)$$

in probability as $n \to \infty$. The theorem is proved.

If $l(\theta, x) = l(\theta)$ is differentiable on θ then $\widetilde{l}_n \overset{a.s.}{\to} 0$ as $n \to \infty$. Really, $\theta_n^* \overset{a.s.}{\to} \theta$, and

$$\widetilde{l}_n = \frac{\ln l(\theta_n^*) - \ln l(\theta)}{(\theta_n^* - \theta)\ln n} \overset{a.s.}{\sim} \frac{l'_\theta(\theta)}{l(\theta)\ln n} \overset{a.s.}{\to} 0.$$

Let $\theta \in (0, 1)$, (1.2) holds and $L_0(n, \theta) \to c(\theta)$ as $n \to \infty$. Then $\alpha(x) = \alpha(x, \theta) \sim x^\theta c^\theta$. For example,

$$p_i(\theta) = \frac{(i - i_0)^{-1/\theta}}{\zeta(1/\theta)}, \ i > i_0,$$

i_0 is integer, $\zeta(z) = \sum_{j=1}^{\infty} j^{-z}$ is Riemann zeta function. In this case $\alpha(\theta, n) = [(n\zeta(1/\theta))^\theta] + i_0$. From SLLN

$$\ln R_n - \theta \ln n - \ln(\Gamma(1-\theta)c^\theta) = \ln n \left(\frac{\ln R_n}{\ln n} - \theta\right) - \ln(\Gamma(1-\theta)c^\theta) \overset{a.s.}{\to} 0.$$

If we use estimator $\theta_n^* = \ln R_n / \ln n$ (it is consistent, Chebunin 2014) then $\ln n(\theta_n^* - \theta)$ tends to some constant a.s. So we need an implicit estimators for asymptotic normality. We construct implicit estimators based on R_n, U_n or $R_{n,k}$. Karlin (1967) proved

$$\mathbf{E}R_n \sim \Gamma(1-\theta)c^\theta n^\theta, \quad \mathbf{Var}R_n \sim \left(2^\theta - 1\right)\Gamma(1-\theta)c^\theta n^\theta, \quad \frac{\mathbf{Var}R_n}{\mathbf{E}R_n} \to 2^\theta - 1,$$

$$\mathbf{E}U_n \sim 2^{\theta-1}\Gamma(1-\theta)c^\theta n^\theta, \quad \mathbf{Var}U_n \sim 4^{\theta-1}\Gamma(1-\theta)c^\theta n^\theta, \quad \frac{\mathbf{Var}U_n}{\mathbf{E}U_n} \to 2^{\theta-1},$$

$$\mathbf{E}R_{n,k} \sim \theta\frac{\Gamma(k-\theta)}{k!}c^\theta n^\theta, \quad \mathbf{Var}R_{n,k} \sim \frac{\theta}{k!}\left(\Gamma(k-\theta) - \frac{2^\theta\Gamma(2k-\theta)}{2^{2k}k!}\right)c^\theta n^\theta,$$

$$\frac{\mathbf{Var}R_{n,k}}{\mathbf{E}R_{n,k}} \to 1 - \frac{2^\theta\Gamma(2k-\theta)}{2^{2k}k!\Gamma(k-\theta)}.$$

LEMMA 1. *If* $\alpha(x) = (cx)^\theta + o(x^{\frac{\theta}{2}})$ *then*

$$\boldsymbol{E}R_n = \Gamma(1-\theta)c^\theta n^\theta + o(n^{\frac{\theta}{2}}), \quad \boldsymbol{E}U_n = 2^{\theta-1}\Gamma(1-\theta)c^\theta n^\theta + o(n^{\frac{\theta}{2}}),$$

$$\boldsymbol{E}R_{n,k} = \theta\frac{\Gamma(k-\theta)}{k!}c^\theta n^\theta + o(n^{\frac{\theta}{2}}).$$

PROOF. The following asymptotics hold under (2) (see Karlin 1967 and Gnedin et al. 2007, Lemma 1)

$$\mathbf{E}(R_n - R_{\Pi(n)}) \to 0, \quad \mathbf{E}(U_n - U_{\Pi(n)}) \to 0, \quad \mathbf{E}(R_{n,k} - R_{\Pi(n),k}) \to 0.$$

We use Karlin (1967) representation, integration by parts and substitution $nt = x$ to get

$$\mathbf{E}R_{\Pi(n)} = \int_0^\infty \left(1-e^{-n/x}\right) d\alpha(x) = \int_0^\infty \alpha(x)nx^{-2}e^{-n/x}dx$$

$$= \int_0^\infty ((cnt)^\theta + o((nt)^{\frac{\theta}{2}}))t^{-2}e^{-1/t}dt = \Gamma(1-\theta)c^\theta n^\theta + o(n^{\frac{\theta}{2}}).$$

Similarly for $\mathbf{E}U_{\Pi(n)}$ and $\mathbf{E}R_{\Pi(n),k}$. The proof is complete.

LEMMA 2. *If* (1.1) *holds then* $\alpha(x) = (cx)^\theta + o(x^{\frac{\theta}{2}})$.

PROOF. For any fixed $\theta \in (0,1)$, convergence $i \to \infty$ takes place if and only if $x \to \infty$. Let us solve equation $c \cdot i^{-1/\theta}(1+\beta(i)) = \frac{1}{x}$ for x large enough, $\beta(i) = o(i^{-\frac{1}{2}})$. We have

$$i = (cx)^\theta(1+\beta(i))^\theta = (cx)^\theta(1+\widetilde{\beta}(i)), \tag{2.3}$$

$\widetilde{\beta}(i) = (1+\beta(i))^\theta - 1 = i^{-1/2}\cdot o(1) = (cx)^{-\theta/2}(1+\widetilde{\beta}(i))^{-1/2}\cdot o(1) = o(x^{-\theta/2})$.

From (2.3),

$$i = (cx)^\theta + o(x^{\frac{\theta}{2}}).$$

The proof is complete.

COROLLARY 1. *If* (1.1) *holds, c is known,* $\frac{dc}{d\theta}$ *exists,* $\theta^*_{n,R}$, $\theta^*_{n,U}$, $\theta^*_{n,k}$ *are the solutions of the equations*

$$R_n = \Gamma(1-\theta)(cn)^\theta, \quad U_n = 2^{\theta-1}\Gamma(1-\theta)(cn)^\theta, \quad R_{n,k} = \theta\frac{\Gamma(k-\theta)}{k!}(cn)^\theta$$

respectively, then

$$\ln n\sqrt{R_n}(\theta^*_{n,R} - \theta) \Rightarrow \boldsymbol{N}_{0,2^\theta-1}, \quad \ln n\sqrt{U_n}(\theta^*_{n,U} - \theta) \Rightarrow \boldsymbol{N}_{0,2^{\theta-1}},$$

$$\ln n\sqrt{R_{n,k}}(\theta^*_{n,k} - \theta) \Rightarrow \boldsymbol{N}_{0,\sigma^2}, \quad \sigma^2 = 1 - \frac{2^\theta\Gamma(2k-\theta)}{2^{2k}k!\Gamma(k-\theta)}.$$

Implicit equations of Corollary 1 rarely can be solved in explicit form. An example of family of distributions with explicit estimator of θ is a family with $c = c_1(\Gamma(1-\theta))^{-1/\theta}$ in (1.1). Here c_1 is a known constant that does not depend on θ. In this case $\theta_n^* = \ln R_n / \ln(c_1 n)$.

Note that one can find similar implicit estimators in more general assumptions than (1.1). For example, one can prove analogs of Theorem 1 and Corollary 1 for function

$$\alpha(x) = \sum_{i=1}^{K} (c_i x)^{\beta_i} + o(x^{\theta/2})$$

with differentiable functions $c_i(\theta) > 0$, $\beta_i(\theta) \in [\theta/2, \theta]$.

3 Explicit Estimators on a Base of Two Statistics

Let the parameter (function) c be unknown. In this case we need two statistics to estimate θ. Some of the following estimators are proposed by Ohannessian and Dahleh (2012). We prove their asymptotical normality. Note that rates of convergence are lower in this case.

Theorem 2. *If* $\frac{\boldsymbol{E}R_{n,1} - \theta \boldsymbol{E}R_n}{\sqrt{\alpha(n)}} \to 0$ *then* $\sqrt{R_n}\left(\frac{R_{n,1}}{R_n} - \theta\right) \Rightarrow \boldsymbol{N}_{0,\sigma_0^2}$,

$$\sigma_0^2 = \theta((9\theta - 1)2^{\theta - 2} + 1 - \theta).$$

Proof. Using SLLN we have

$$\begin{aligned}\sqrt{R_n}\left(\frac{R_{n,1}}{R_n} - \theta\right) &= \frac{R_{n,1} - \theta R_n}{\sqrt{R_n}} \overset{a.s.}{\sim} \frac{R_{n,1} - \theta R_n}{\sqrt{\Gamma(1-\theta)\alpha(n)}} \\ &\overset{a.s.}{\sim} \frac{R_{n,1} - \mathbf{E}R_{n,1} - \theta(R_n - \mathbf{E}R_n)}{\sqrt{\Gamma(1-\theta)\alpha(n)}} \\ &= \frac{1}{\sqrt{\Gamma(1-\theta)}}\left(\frac{R_{n,1} - \mathbf{E}R_{n,1}}{\sqrt{\alpha(n)}} - \theta\frac{R_n - \mathbf{E}R_n}{\sqrt{\alpha(n)}}\right).\end{aligned}$$

Then we calculate limiting variance using Corollary 3. The proof is complete.

Note that $\sigma_0^2 < 4$ for $\theta \in (0,1)$.

Theorem 3. *If* $\frac{(k-\theta)\boldsymbol{E}R_{n,k} - (k+1)\boldsymbol{E}R_{n,k+1}}{\sqrt{\alpha(n)}} \to 0$ *then*

$$\sqrt{R_{n,k}}\left(\frac{kR_{n,k} - (k+1)R_{n,k+1}}{R_{n,k}} - \theta\right) \Rightarrow \boldsymbol{N}_{0,\sigma_k^2},$$

$$\sigma_k^2 = (k-\theta)(2k+1-\theta) - \frac{(2k-\theta+\theta^2)}{k 2^{2k+2-\theta} \mathrm{B}(k-\theta, k)},$$

B *is a Beta function.*

PROOF. Using SLLN we have

$$\sqrt{R_{n,k}}\left(\frac{kR_{n,k}-(k+1)R_{n,k+1}}{R_{n,k}}-\theta\right)=\frac{(k-\theta)R_{n,k}-(k+1)R_{n,k+1}}{\sqrt{R_{n,k}}}$$

$$\overset{a.s.}{\sim}\frac{(k-\theta)(R_{n,k}-\mathbf{E}R_{n,k})-(k+1)(R_{n,k+1}-\mathbf{E}R_{n,k+1})}{\sqrt{\theta\frac{\Gamma(k-\theta)}{k!}\alpha(n)}}$$

$$=\frac{1}{\sqrt{\theta\frac{\Gamma(k-\theta)}{k!}}}\left((k-\theta)\frac{R_{n,k}-\mathbf{E}R_{n,k}}{\sqrt{\alpha(n)}}-(k+1)\frac{R_{n,k+1}-\mathbf{E}R_{n,k+1}}{\sqrt{\alpha(n)}}\right).$$

Then we calculate limiting variance on the base of Theorem 5 in Karlin (1967). The proof is complete.

From Lemmas 1 and 2 we obtain the following corollary.

COROLLARY 2. *Assumptions of* Theorems 2 *and* 3 *are held under* (1.1).

Acknowledgments. Our research was partially supported by RFBR grant 17-01-00683 and by the program of fundamental scientific researches of the SB RAS No. I.1.3., project No. 0314-2016-0008.

References

BAHADUR, R.R. (1960). On the number of distinct values in a large sample from an infinite discrete distribution. *Proceedings of the National Institute of Sciences of India* **26A, Supp II**, 67–75.

BARBOUR, A.D. (2009). Univariate approximations in the infinite occupancy scheme. *Alea* **6**, 415–433.

BARBOUR, A.D. and GNEDIN, A.V. (2009). Small counts in the infinite occupancy scheme. Electronic. *J. Probab.* **14**, 365–384.

BEN-HAMOU, A., BOUCHERON, S. and GASSIAT, E. (2016). Pattern coding meets censoring: (almost) adaptive coding on countable alphabets. arXiv:1608.08367.

BEN-HAMOU, A., BOUCHERON, S. and OHANNESSIAN, M.I. (2017). Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications. *Bernoulli* **23**, 249–287.

BOGACHEV, L.V., GNEDIN, A.V. and YAKUBOVICH, Y.V. (2008). On the variance of the number of occupied boxes. *Adv. Appl. Math.* **40**, 401–432.

BOONTA, S. and NEAMMANEE, K. (2007). Bounds on random infinite urn model. *Bull. Malays. Math. Sci. Soc. Second Series* **30.2**, 121–128.

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

CHEBUNIN, M. and KOVALEVSKII, A. (2016). Functional central limit theorems for certain statistics in an infinite urn scheme. *Statist. Probab. Lett.* **119**, 344–348.

DURIEU, O. and WANG, Y. (2016). From infinite urn schemes to decompositions of self-similar Gaussian processes. *Electron. J. Probab.* **21**, 43.

DUTKO, M. (1989). Central limit theorems for infinite urn models. *Ann. Probab.* **17**, 1255–1263.

GNEDIN, A., HANSEN, B. and PITMAN, J. (2007). Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Probab. Surv.* **4**, 146–171.

GRUBEL, R. and HITCZENKO, P. (2009). Gaps in discrete random samples. *J. Appl. Probab.* **46**, 1038–1051.

HEAPS, H.S. (1978). *Information retrieval, computational and theoretical aspects.* Academic Press.

HERDAN, G. (1960). *Type-token mathematics.* The Hague, Mouton.

HWANG, H.-K. and JANSON, S. (2008). Local limit theorems for finite and infinite urn models. *Ann. Probab.* **36**, 992–1022.

KARLIN, S. (1967). Central limit theorems for certain infinite urn schemes. *J. Math. Mech.* **17**, 373–401.

KEY, E.S. (1992). Rare Numbers. *J. Theor. Probab.* **5**, 375–389.

KEY, E.S. (1996). Divergence rates for the number of rare numbers. *J. Theor. Probab.* **9**, 413–428.

KHMALADZE, E.V. (2011). Convergence properties in certain occupancy problems including the Karlin-Rouault law. *J. Appl. Probab.* **48**, 1095–1113.

MANDELBROT, B. (1965). Information theory and psycholinguistics. In *Scientific psychology. Basic Books*, (B.B. Wolman and E. Nagel, eds.)

MURATOV, A. and ZUYEV, S. (2016). Bit flipping and time to recover. *J. Appl. Probab.* **53**, 650–666.

NICHOLLS, P.T. (1987). Estimation of Zipf parameters. *J. Am. Soc. Inf. Sci.* **38**, 443–445.

OHANNESSIAN, M.I. and DAHLEH, M.A. (2012). Rare probability estimation under regularly varying heavy tails. In *Proceedings of the 25th Annual Conference on Learning Theory PMLR*, pp. 23:21.1–21.24.

PETERSEN, A.M., TENENBAUM, J.N., HAVLIN, S., STANLEY, H.E. and PERC, M. (2012). Languages cool as they expand: allometric scaling and the decreasing need for new words. Scientific Reports 2. Article No 943.

ZAKREVSKAYA, N.S. and KOVALEVSKII, A.P. (2001). One-parameter probabilistic models of text statistics. *Sib. Zh. Ind. Mat.* **4:2**, 142–153. (in Russian).

ZIPF, G.K. (1949). *Human behavior and the principle of least effort.* University Press, Cambridge.

Appendix: Functional Central Limit Theorem

Let for $t \in [0,1],\ k \geq 1$

$$Y^*_{n,k}(t) = \frac{R^*_{[nt],k} - \mathbf{E}R^*_{[nt],k}}{(\alpha(n))^{1/2}}, \qquad Y_{n,k}(t) = \frac{R_{[nt],k} - \mathbf{E}R_{[nt],k}}{(\alpha(n))^{1/2}}.$$

THEOREM 4. *Let us assume that* (1.2) *holds,* $\nu \geq 1$ *is integer. Then random process* $\big((Y^*_{n,1}(t), Y_{n,1}(t), \ldots, Y_{n,\nu}(t)),\ 0 \leq t \leq 1\big)$ *converges weakly* *in the uniform metrics in $D(0,1)$ to $(\nu+1)$-dimensional Gaussian process with continuous sample paths, zero expectation and covariance function $(c_{ij}(\tau,t))_{i,j=0}^{\nu}$,*

$$c_{ij}(\tau,t) = \frac{\theta\tau^i (t-\tau)^{j-i} t^{\theta-j}\Gamma(j-\theta)}{i!(j-i)!} - \frac{\theta\tau^i t^j (t+\tau)^{\theta-i-j}\Gamma(i+j-\theta)}{i!j!}$$
$$\text{for } 1 \le i \le j,\ \tau \le t,$$
$$c_{ij}(\tau,t) = -\frac{\theta\tau^i t^j (t+\tau)^{\theta-i-j}\Gamma(i+j-\theta)}{i!j!} \quad \text{for } i > j \ge 1,\ \tau \le t,$$
$$c_{00}(\tau,t) = \left((t+\tau)^\theta - t^\theta\right)\Gamma(1-\theta) \quad \text{for } \tau \le t,$$
$$c_{i0}(\tau,t) = -\frac{\theta\tau^i (t+\tau)^{\theta-i}\Gamma(i-\theta)}{i!} \quad \text{for } i > 0,\ \tau \le t,$$
$$c_{0j}(\tau,t) = \frac{\theta((t-\tau)^j t^{\theta-j} - t^j (t+\tau)^{\theta-j})\Gamma(j-\theta)}{j!} \quad \text{for } j > 0,\ \tau \le t,$$

$c_{ji}(t,\tau) = c_{ij}(\tau,t)$.

Proof.

Theorem 3 by Chebunin and Kovalevskii (2016) states weak convergence of vector random process $\big((Y^*_{n,1}(t),\ldots,Y^*_{n,\nu}(t)),\ 0 \le t \le 1\big)$ in the uniform metrics in $D(0,1)$ to $(\nu+1)$-dimensional Gaussian process with continuous sample paths, zero expectation and covariance function $(c^*_{ij}(\tau,t))_{i,j=0}^{\nu}$.

The main focus of this paper was to prove tightness of components $(Y^*_{n,i}(t),\ 0 \le t \le 1)$ by Poissonization and construction of an appropriate inequality for covariances.

As $Y_{n,i}(t) = Y^*_{n_i}(t) - Y^*_{n,i-1}(t)$, we state tightness of components $(Y_{n,i},\ 0 \le t \le 1)$ and calculate $c_{ij}(\tau,t)$ by formulas

$$c_{ij}(\tau,t) = c^*_{ij}(\tau,t) - c^*_{i+1,j}(\tau,t) - c^*_{i,j+1}(\tau,t) + c^*_{i+1,j+1}(\tau,t),$$

$$c_{0j}(\tau,t) = c^*_{1j}(\tau,t) - c^*_{1,j+1}(\tau,t), \quad c_{i0}(\tau,t) = c^*_{i1}(\tau,t) - c^*_{i+1,1}(\tau,t).$$

The proof is complete. □

The limiting $(\nu+1)$-dimensional Gaussian process is self-similar with Hurst parameter $H = \theta/2 < 1/2$. Its first component coincides in distribution with the first component of the limiting process in Theorem 1 in Durieu and Wang (2016).

We need some specific corollary to calculate limiting variance in Theorem 2.

Corollary 3. *In assumptions of* Theorem 4*, random vector* $((Y^*_{n,1}(1), Y_{n,1}(1))$ *converges weakly to a normal one with zero mean and covariance matrix*

$$\Gamma(1-\theta)\begin{pmatrix} 2^\theta - 1 & -\theta 2^{\theta-1} \\ -\theta 2^{\theta-1} & \theta(1-2^{\theta-2}(1-\theta)) \end{pmatrix}.$$

Mikhail Chebunin
Sobolev Institute of Mathematics,
Novosibirsk, Russia
Novosibirsk State University,
Novosibirsk, Russia
E-mail: chebuninmikhail@gmail.com

Artyom Kovalevskii
Novosibirsk State Technical
University, Novosibirsk, Russia
Novosibirsk State University,
Novosibirsk, Russia
E-mail: kovalevskiii@gmail.com

Paper received: 14 June 2017.