S©MR ISSN 1813-3304

СИБИРСКИЕ ЭЛЕКТРОННЫЕ
МАТЕМАТИЧЕСКИЕ ИЗВЕСТИЯ

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

Том 16, стр. 1822–1832 (2019) УДК 519.233
DOI 10.33048/semi.2019.16.129 MSC 62F03

A STATISTICAL TEST FOR THE ZIPF'S LAW BY DEVIATIONS FROM THE HEAPS' LAW

M.G. CHEBUNIN, A.P. KOVALEVSKII

Abstract. We explore a probabilistic model of an artistic text: words of the text are chosen independently of each other in accordance with a discrete probability distribution on an infinite dictionary. The words are enumerated 1, 2, ..., and the probability of appearing the i'th word is asymptotically a power function. Bahadur proved that in this case the number of different words as a function of the length of the text, again, asymptotically behaves like a power function. On the other hand, in the applied statistics community there are statements known as the Zipf's and Heaps' laws that are supported by empirical observations. We highlight the links between Bahadur results and Zipf's/Heaps' laws, and introduce and analyse a corresponding statistical test.

Keywords: Zipf's law, Heaps' law, weak convergence.

1. Introduction

There is a countably infinite dictionary where the words are numbered 1, 2, ... Words are chosen one-by-one independently of each other and accordingly to a discrete probability distribution on the positive integers that is equivalent to a power law distribution

$$p_i \sim ci^{-1/\theta}, \quad 0 < \theta < 1, \quad c > 0. \tag{1}$$

We assume further that the sequence $\{p_n\}$ is decreasing, $p_{n+1} \leq p_n$ for all n. Hereinafter, for two positive sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \sim b_n$ if $a_n/b_n \to 1$, as $n \to \infty$.

Chebunin, M.G., Kovalevskii, A.P., A statistical test for the Zipf's law by deviations from the Heaps' law.

The work is supported by RFBR (grant 17-01-00683) and by the program of fundamental scientific researches of the SB RAS № I.1.3., project № 0314-2019-0008.

Received September, 24, 2019, published December, 4, 2019.

Let R_n be the number of different words in the text of length n. It may be represented as $R_n = \sum_{j\geq 1} I_j$ where I_j is the indicator function, corresponding to the event where the j-th word of the dictionary is present in the text.

Bahadur (1960) proved that under (1) holds

$$\mathbf{E}R_n \sim C_1 n^\theta, \tag{2}$$

where $C_1 = c^\theta \Gamma(1-\theta)$ and $\Gamma(x) = \int_0^\infty y^{x-1}e^{-y}\,dy$ is the Euler gamma function. Bahadur also proved convergence in probability $R_n/\mathbf{E}R_n \xrightarrow{p} 1$.

Karlin (1967) made next important steps. He proved that $R_n/\mathbf{E}R_n \xrightarrow{a.s.} 1$, which, combined with (2), is equivalent to

$$R_n \sim C_1 n^\theta \text{ a.s.} \tag{3}$$

In fact, Karlin studied a more general problem. He assumed that p_i satisfy the following regular variation property:

$$\alpha(x) = \max\{j|\ p_j \geq 1/x\} = x^\theta L(x),\ \text{for } \theta \in [0,1], \tag{4}$$

where $L(x)$ is a function slowly varying at infinity. He considered texts of a random length using an independent rate-1 Poisson process $\Pi(t)$: the length of the text grows in time and follows a Poisson distribution $\Pi(t)$ at time $t > 0$. Let $R_{n,k}$ be the number of words with k occurrences. Using this Poissanization, Karlin proved the Central Limit Theorems for R_n and $R_{n,k}$ under condition (4) for $\theta > 0$.

Within the last decades, the theory of infinite urn schemes has been developed in several directions.

First, Key (1992, 1996) studied general properties of the number of words $R_{n,1}$ that appears only once in the text. Muratov and Zuyev (2016) studied properties of $R_{n,1}$ using their Markov chain representation.

Secondly, Ben-Hamou, Boucheron and Ohannessian (2017) showed that R_n and $R_{n,k}$ satisfy Bernstein-type concentration inequalities, without assuming extra condition (4). The variance factors in these concentration inequalities are shown to be tight in the regular case. Decrouez, Grabchak and Paris (2018) gave upper and lower bounds for the expected occupancy counts $\mathbf{E}R_{n,k}$ in terms of the function $\alpha(x)$. If $\alpha(x)$ is bounded above and below by regularly varying functions, then their general results lead to an optimal-rate control of the expected occupancy counts.

Thirdly, the case $\theta = 0$ was studied. Dutko (1989) proved the CLT for R_n under condition $\mathbf{Var}R_n \to \infty$, as $n \to \infty$. This is always true for $\theta > 0$ and may also hold for $\theta = 0$ in a particular case. Barbour and Gnedin (2009) proved the CLT for $R_{n,k}$ under condition $\mathbf{Var}R_{n,k} \to \infty$, as $n \to \infty$, and Hwang and Janson (2008) proved the Local CLT.

Fourthly, in the regular case, there are known functional versions of the CLT. Chebunin and Kovalevskii (2016) proved the Functional Central Limit Theorem for R_n and $R_{n,k}$ under condition (4) for $\theta \in (0,1)$. So, the process

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

converges weakly to a centered Gaussian process Z_θ with continuous a.s. sample paths and covariance function of the form

$$K(s,t) = (s+t)^\theta - \max(s^\theta, t^\theta)$$

(note that $\mathbf{Var}R_n \sim (2^\theta - 1)\mathbf{E}R_n$).

In the case $\theta = 1$, Chebunin and Kovalevskii (2016) and Chebunin (2017) proved the FCLT for R_n and $R_{n,k}$, correspondingly. Durieu and Wang (2016) introduced a natural randomization of R_n for parameter $\theta \in (0, 1)$, where each indicator function is multiplied independently by a random variable taking values ± 1 with equal probabilities, and proved the FCLT for this case. Durieu, Samorodnitsky and Wang (2019) investigated the odd-occupancy process for the randomized Karlin model with parameter $\theta \in (0, 1)$, where each indicator function is multiplied independently by a random variable with the heavy-tailed distribution. They proved that this process converges to a stable process with stationary increments.

Gnedin, Hansen and Pitman (2007) analysed further properties of R_n, $R_{n,k}$ and of their Poissonizations. Barbour (2009) proposed approximations for R_n and $R_{n,k}$ within the family of translated Poisson distributions.

Various estimators of the parameter θ have been obtained and analysed by Nicholls (1978), Zakrevskaya and Kovalevskii (2001, 2019), Guillou and Hall (2002), Ohannessian and Dahleh (2012), Chebunin (2014), Chebunin and Kovalevskii (2018).

In the applied statistics, relations (1) and (3) were observed empirically in the analysis of artistic texts. Linguists call them the Zipf's law and the Heaps' law respectively. The Zipf's law (Zipf, 1936) states the decrease in the frequencies of words depending on the rank in accordance with a power function.

The content of the Heaps' law was initially proposed by Herdan (1960). This law was then popularized by Heaps (1978). The Heaps' law states that dependence of the number of different words from the text length is characterised by a power function.

Links between the Heaps' and the Zipf's laws have been studied (empirically and in other contexts) by van Leijenhorst and van der Weide (2005), Serrano et al. (2009), Bernhardsson et al. (2009), Eliazar (2011), Baeza-Yates & Navarro (2013), etc.

We could not find in the literature any mathematically correct statistical goodness-of-fit test for the Zipf's law. Altmann and Gerlach (2016) emphasize incorrectness of a number of statistical tests proposed earlier.

Note that analysis of very long texts and text sequences shows significant deviations of words frequencies from the Zipf model (see, for example, Petersen et al., (2012)). Gerlach and Altmann (2013) proposed a modified model for explanations of these deviations.

The present paper proposes a new theoretically supported test for the Zipf's law. We introduce a new class of estimates that is based on the sequence $(R_1, \dots, R_n)$. We define an empirical process and prove its weak convergence to a centered Gaussian process. We calculate the covariance function of this limiting process. Then we construct a test of the omega-squared type. Calculation of the limiting distribution of the test statistics is based on the corresponding classical result of Smirnov (1937). One can calculate this distribution using the results of Deheuvels and Martynov (1996).

The rest of the paper is organised as follows. We propose an estimator for parameter θ and state its properties in Section 2. Then we propose a test for known θ and state its properties in Section 3. Then a new test for unknown θ (that is, the test for the Zipf's law) follows in Section 4. The proofs of the formulated results are given in Section 5.

2. Parameter's estimation

From (3), we have $\log R_n \sim \theta \log n$ a.s. Therefore, we may propose the following estimator for parameter θ:

$$\widehat{\theta} = \int_0^1 \log^+ R_{[nt]}\, dA(t),$$

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

$$A(0) = A(1) = 0, \quad \lim_{x \downarrow 0} \log x \int_0^x |dA(t)| = 0, \quad \int_0^1 \log t\, dA(t) = 1. \tag{5}$$

Theorem 1. *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.*

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

Theorem 2. *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*

$$\sqrt{\mathbf{E}R_n}(\widehat{\theta} - \theta) - \int_0^1 t^{-\theta} Z_n(t)\, dA(t) \to_p 0.$$

From Theorem 2, it follows that $\widehat{\theta}$ converges to θ with rate $(\mathbf{E}R_n)^{-1/2}$, and $\sqrt{\mathbf{E}R_n}(\widehat{\theta} - \theta)$ converges weakly to the normal random variable $\int_0^1 t^{-\theta} Z_\theta(t)\, dA(t)$ with variance $\int_0^1 \int_0^1 (st)^{-\theta} K(s, t)\, dA(s)\, dA(t)$.

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

$$\widehat{\theta} = \log_2(R_n / R_{[n/2]}), \quad n \ge 2.$$

Note that, in this example, for any function g on $[0, 1]$,

$$\int_0^1 g(t)\, dA(t) = \frac{g(1) - g(1/2)}{\log 2}.$$

3. Test for a known rate

Let $0 < \theta < 1$ be known. We introduce an *empirical bridge* Z_n^0 (Kovalevskii and Shatalin, 2015, 2016) as follows.

$$Z_n^0(k/n) = \left(R_k - (k/n)^\theta R_n\right) / \sqrt{R_n},$$

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

$$Z_n^0\left(\frac{k}{n} + u\right) = Z_n^0(k/n) + nu\left(Z_n^0((k+1)/n) - Z_n^0(k/n)\right).$$

Theorem 3. *Under the assumptions of Theorem 2,*

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

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 the FCLT of Chebunin & Kovalevskii (2016), we have

Corollary 1. *Under the assumptions of Theorem 2, Z_n^0 converges weakly in $C(0,1)$ to a Gaussian process Z_θ^0 that can be represented as $Z_\theta^0(t) = Z_\theta(t) - t^\theta Z_\theta(1)$, $0 \le t \le 1$. Its correlation function is given by*

$$K^0(s,t) = \mathbf{E} Z_\theta^0(s) Z_\theta^0(t) = K(s,t) - s^\theta K(1,t) - t^\theta K(s,1) + s^\theta t^\theta K(1,1).$$

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

Let $W_n^2 = \int\limits_0^1 \left(Z_n^0(t)\right)^2 dt$. It is equal to

$$W_n^2 = \frac{1}{3n} \sum_{k=1}^{n-1} Z_n^0\left(\frac{k}{n}\right)\left(2Z_n^0\left(\frac{k}{n}\right) + Z_n^0\left(\frac{k+1}{n}\right)\right). \tag{6}$$

Then W_n^2 converges weakly to $W_\theta^2 = \int\limits_0^1 \left(Z_\theta^0(t)\right)^2 dt$.

So the test rejects the basic hypothesis if $W_n^2 \ge C$. 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 a concrete value of W_n^2 for observations under consideration.

One can estimate F_θ by simulations or find it explicitely using the Smirnov's 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

$$F_\theta(x) = 1 + \frac{1}{\pi} \sum_{k=1}^\infty (-1)^k \int_{\lambda_{2k-1}}^{\lambda_{2k}} \frac{e^{-\lambda x/2}}{\sqrt{-D(\lambda)}} \cdot \frac{d\lambda}{\lambda},\ x > 0, \tag{7}$$

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

The integrals in the RHS of (7) must tend to 0 monotonically as $k \to \infty$, and λ_k^{-1} are the eigenvalues of kernel K^0 (see Martynov (1973), Chapter 3).

4. Test for an unknown rate

Let us introduce the process $\widehat{Z}_n$:

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

$0 \le k \le n$. As for Z_n^0, let for $0 \le u < 1/n$ and $0 \le k \le n-1$

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

Theorem 4. *Under assumptions of Theorem 2, $\widehat{Z}_n$ converges weakly to $\widehat{Z}_\theta$ as $n \to \infty$, where*

$$\widehat{Z}_\theta(t) = Z_\theta^0(t) - t^\theta \log t \int_0^1 u^{-\theta} Z_\theta(u)\, dA(u).$$

Corollary 2. *Assume the conditions of Theorem 2 to hold. Let* $\widehat{W}_n^2 = \int\limits_0^1 \left(\widehat{Z}_n(t)\right)^2 dt$. *Then* $\widehat{W}_n^2$ *converges weakly to* $\widehat{W}_\theta^2 = \int\limits_0^1 \left(\widehat{Z}_\theta(t)\right)^2 dt$.

Similarly to (6), $\widehat{W}_n^2$ has the following representation

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

The p-value of the goodness-of fit test is $1-\widehat{F}_\theta(\widehat{W}_{n,obs}^2)$. Here $\widehat{F}_\theta$ is the cumulative distribution function of $\widehat{W}_\theta^2$, and $\widehat{W}_{n,obs}^2$ is the observed value of $\widehat{W}_n^2$. Further, the function $\widehat{F}_\theta$ can be found using the approach from Section 3, with replacing λ_k by $\widehat{\lambda}_k$ in the Smirnov's formula, and $\widehat{\lambda}_k$ are the eigenvalues of the kernel $\widehat{K}(s,t) = \mathbf{E}\widehat{Z}_\theta(s)\widehat{Z}_\theta(t)$.

5. The proofs of formulated results

Proof of Theorem 1

Since $p_i i^{1/\theta}$ is a slowly varying function as $i \to \infty$, we have $\alpha(x) = x^\theta L(x,\theta)$, $L(x,\theta)$ is a slowly varying function as $x \to \infty$ (Karlin, 1967).

Let $\delta_n = 1/\sqrt{n}$. Then

$$\left|\int_0^{\delta_n} \log^+ R_{[nt]}\, dA(t)\right| \le_{a.s.} \int_0^{1/\sqrt{n}} \log^+ nt\ |dA(t)|$$

$$= \int_{1/n}^{1/\sqrt{n}} \log nt\ |dA(t)| \le \int_{1/n}^{1/\sqrt{n}} \log n\ |dA(t)| + \int_{1/n}^{1/\sqrt{n}} \log t\ |dA(t)|$$

$$= -2\log\frac{1}{\sqrt{n}} \int_{1/n}^{1/\sqrt{n}} |dA(t)| + o(1).$$

As $\int_{1/n}^{1/\sqrt{n}} |dA(t)| \le \int_0^{1/\sqrt{n}} |dA(t)|$, from (5) the RHS tends to 0.

The rest of the integral is

$$\int_{\delta_n}^1 \log R_{[nt]}\, dA(t) = \int_{\delta_n}^1 \log\frac{R_{[nt]}}{\mathbf{E}R_{[nt]}}\, dA(t) + \int_{\delta_n}^1 \log \mathbf{E}R_{[nt]}\, dA(t).$$

We prove a.s. convergence to 0 of the first integral in RHS, and then a.s. convergence to θ of the second one.

By the SLLN, $\log(R_j/\mathbf{E}R_j) \to 0$ a.s. as $j \to \infty$. Therefore, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbf{P}\left(\sup_{j\ge n\delta_n}\left|\log\left(\frac{R_j}{\mathbf{E}R_j}\right)\right| \ge \varepsilon\right) = 0.$$

Therefore

$$\mathbf{P}\left(\sup_{k\ge n}\left|\int_{\delta_n}^1 \log\frac{R_{[kt]}}{\mathbf{E}R_{[kt]}}\, dA(t)\right| \ge \varepsilon\right) \le \mathbf{P}\left(\int_{\delta_n}^1 \sup_{k\ge n}\left|\log\frac{R_{[kt]}}{\mathbf{E}R_{[kt]}}\right| |dA(t)| \ge \varepsilon\right)$$

$$= \mathbf{P}\left(\sup_{k\ge n\delta_n}\left|\log\frac{R_k}{\mathbf{E}R_k}\right| \ge \frac{\varepsilon}{\int_{\delta_n}^1 |dA(t)|}\right) \to 0 \quad \text{as} \quad n\to\infty,$$

that is, $\int_{\delta_n}^1 \log \frac{R_{[nt]}}{\mathbf{E}R_{[nt]}}\, dA(t) \to 0$ a.s.

We have $\mathbf{E}R_j = j^\theta L(j)$, where $L(j)$ is a slowly varying function (Karlin, 1967). So, uniformly in $t \geq \delta_n > 0$,

$$\log \frac{\mathbf{E}R_{[nt]}}{(nt)^\theta L(nt)} \to 0.$$

Therefore

$$\int_{\delta_n}^1 \log \frac{\mathbf{E}R_{[nt]}}{(nt)^\theta L(nt)}\, dA(t) \to 0.$$

However, $L(nt) = (nt)^{o(1)}$ for $t \geq \delta_n$, so from (5) we have

$$\int_{\delta_n}^1 \log((nt)^\theta L(nt))\, dA(t) = \int_{\delta_n}^1 \log((nt)^{\theta+o(1)})\, dA(t) = (\theta + o(1)) \int_{\delta_n}^1 \log(nt)\, dA(t)$$

$$= (\theta + o(1)) \int_0^1 \log(nt)\, dA(t) - (\theta + o(1)) \int_0^{\delta_n} \log(nt)\, dA(t) = \theta + o(1).$$

Then

$$\int_{\delta_n}^1 \log \mathbf{E}R_{[nt]}\, dA(t) \to \theta,$$

and $\widehat{\theta} \to \theta$ a.s. The proof is complete.

Proof of Theorem 2

Since $p_i = ci^{-1/\theta}(1 + o(i^{-1/2}))$, we have

$$\mathbf{E}R_n = C_1 n^\theta + o(n^{\frac{\theta}{2}}) \tag{8}$$

(Chebunin and Kovalevskii, 2018, Lemma 1 and 2). Recall that

$$Z_n(t) = \frac{R_{[nt]} - \mathbf{E}R_{[nt]}}{\sqrt{\mathbf{E}R_n}}.$$

Let

$$Z_n^*(t) = \frac{R_{[nt]} - \mathbf{E}R_{[nt]}}{\mathbf{E}R_{[nt]}}.$$

Then

$$\sqrt{\mathbf{E}R_n} \left(\int_0^1 \log R_{[nt]}\, dA(t) - \theta \right) - \int_0^1 t^{-\theta} Z_n(t)\, dA(t)$$

$$= \sqrt{\mathbf{E}R_n} \left(\int_0^1 \left(\log C_1 (nt)^\theta + \log \frac{\mathbf{E}R_{[nt]}}{C_1(nt)^\theta} + \log(1 + Z_n^*(t)) - t^{-\theta} \frac{Z_n(t)}{\sqrt{\mathbf{E}R_n}} \right) dA(t) - \theta \right)$$

$$= \sqrt{\mathbf{E}R_n} \int_0^1 \left(\log \frac{\mathbf{E}R_{[nt]}}{C_1(nt)^\theta} + \log(1 + Z_n^*(t)) - Z_n^*(t) + Z_n^*(t) - t^{-\theta} \frac{Z_n(t)}{\sqrt{\mathbf{E}R_n}} \right) dA(t).$$

For any $t \in [\delta, 1]$, $Z_n^*(t) \to 0$ a.s. , so $\log(1 + Z_n^*(t)) - Z_n^*(t) \sim -(Z_n^*(t))^2/2$ a.s. We have $\mathbf{E}R_n = n^\theta L(n)$, so, for any $t \in [\delta, 1]$,

$$\frac{(\mathbf{E}R_n)^{\frac{3}{2}}}{(\mathbf{E}R_{[nt]})^2} = \frac{(n^\theta L(n))^{\frac{3}{2}}}{((nt)^\theta L(nt))^2} = O\left(\frac{1}{n^{\frac{\theta}{2}} \sqrt{L(n)}} \right), \quad \frac{t^\theta \mathbf{E}R_n}{\mathbf{E}R_{[nt]}} = \frac{L(n)}{L(nt)} = 1 + o(1).$$

Note that

$$\sqrt{\mathbf{E}R_n} \int_0^1 (Z_n^*(t))^2\, dA(t) = \int_0^1 \frac{(\mathbf{E}R_n)^{\frac{3}{2}}}{(\mathbf{E}R_{[nt]})^2} (Z_n(t))^2\, dA(t)$$

$$= o(n^{-\theta/4}) \int_0^1 (Z_n(t))^2 \, dA(t) \to_p 0,$$

so

$$\sqrt{\mathbf{E}R_n} \int_0^1 \left(Z_n^*(t) - t^{-\theta} \frac{Z_n(t)}{\sqrt{\mathbf{E}R_n}} \right) dA(t) = \int_0^1 \left(\frac{t^\theta \mathbf{E}R_n}{\mathbf{E}R_{[nt]}} - 1 \right) t^{-\theta} Z_n(t) \, dA(t)$$

$$= o(1) \int_0^1 t^{-\theta} Z_n(t) \, dA(t) \to 0.$$

Since (8), we have

$$\sqrt{\mathbf{E}R_n} \int_0^1 \log \frac{\mathbf{E}R_{[nt]}}{C_1 (nt)^\theta} \, dA(t) = \sqrt{\mathbf{E}R_n} \int_0^1 \log(1+o(n^{-\frac{\theta}{2}})) \, dA(t) = \int_0^1 o(1) \, dA(t) \to 0.$$

The proof is complete.

Proof of Theorem 3

Let $t \in [0,1)$, $k = [nt]$, 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 $Z_n^0(t)$,

$$\frac{R_k - \left(\frac{k+1}{n}\right)^\theta R_n}{\sqrt{R_n}} \le Z_n^0(t) \le \frac{R_{k+1} - \left(\frac{k}{n}\right)^\theta R_n}{\sqrt{R_n}},$$

so

$$\left| Z_n^0(t) - \frac{R_{[nt]} - t^\theta R_n}{\sqrt{R_n}} \right| \le \frac{R_{k+1} - R_k + \frac{1}{n^\theta} f_\theta(k) R_n}{\sqrt{R_n}}$$

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

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

The proof is complete.

Proof of Theorem 4

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 3.

By the definition,

$$\widehat{Z}_n(k/n) = Z_n^0(k/n) + \sqrt{R_n} \left((k/n)^\theta - (k/n)^{\widehat{\theta}} \right),$$

$$\widehat{Z}_n(t) = Z_n^0(t) + \sqrt{R_n} \left((k/n)^\theta - (k/n)^{\widehat{\theta}} \right)$$

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

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

$$\left| \widehat{Z}_n(t) - Z_n^0(t) + \sqrt{R_n} \left(t^{\widehat{\theta}} - t^\theta \right) \right|$$

$$= \left| \widehat{Z}_n(t) - Z_n^0(t) + \sqrt{R_n} \left(\left(\frac{k}{n} + u \right)^{\widehat{\theta}} - \left(\frac{k}{n} + u \right)^\theta \right) \right|$$

$$\leq 2\sqrt{R_n}\left(f_\theta(k)/n^\theta + f_{\widehat{\theta}}(k)/n^{\widehat{\theta}}\right) \leq 2\sqrt{R_n}\left(1/n^\theta + 1/n^{\widehat{\theta}}\right) \to 0$$

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

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

$$t^{\widehat{\theta}} - t^\theta = t^\theta\left(e^{(\widehat{\theta}-\theta)\log t} - 1\right)$$

$$= (\widehat{\theta}-\theta)t^\theta \log t + t^\theta \sum_{k\geq 2} \frac{((\widehat{\theta}-\theta)\log t)^k}{k!}$$

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

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

$$= (\widehat{\theta}-\theta)t^\theta \log t(1+o(1))$$

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

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

From Theorems 2 and 3, we have joint weak convergence of $(Z_n^0,\ \sqrt{R_n}(\widehat{\theta}-\theta))$ to $(Z_\theta^0,\ \int_0^1 u^{-\theta} Z_\theta(u)\, dA(u))$. So, $\widehat{Z}_n$ converges weakly to $\widehat{Z}_\theta$,

$$\widehat{Z}_\theta(t) = Z_\theta^0(t) - t^\theta \log t \int_0^1 u^{-\theta} Z_\theta(u)\, dA(u).$$

The proof is complete.

Acknowledgements

The research was supported by RFBR grant 17-01-00683 and by the program of fundamental scientific researches of the SB RAS № I.1.3., project № 0314-2019-0008. The authors like to thank Sergey Foss for helpful and constructive comments and suggestions.

References

[1] E.G. Altmann, M. Gerlach, *Statistical laws in linguistics*. In: M. Degli Esposti et al. (eds.), *Creativity and Universality in Language*, Lecture Notes in Morphogenesis, 2016.

[2] R. Baeza-Yates, G. Navarro, *Block Addressing Indices for Approximate Text Retrieval*, J. Am. Soc. Inf. Sci. **51**, 69 (2000).

[3] R.R. Bahadur, *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 (1960), 67–75. MR0137256

[4] A.D. Barbour, *Univariate approximations in the infinite occupancy scheme*, Alea **6** (2009), 415–433. MR2576025

[5] A.D. Barbour, A.V. Gnedin, *Small counts in the infinite occupancy scheme*, Electronic Journal of Probability, **14**, Paper no. 13 (2009), 365–384. MR2480545

[6] 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**, Number 1 (2017), 249–287. MR3556773

[7] S. Bernhardsson, L.E. Correa da Rocha, P. Minnhagen, *The Meta Book and Size-Dependent Properties of Written Language*, New J. Phys. **11**, 123015 (2009).

[8] 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 (in Russian). MR3364413
[9] M.G. Chebunin, *Functional central limit theorem in an infinite urn scheme for distributions with superheavy tails*, Sib. Elektron. Mat. Izv., **14** (2017), 1289–1298. MR3744074
[10] M. Chebunin, A. Kovalevskii, *Functional central limit theorems for certain statistics in an infinite urn scheme*, Statistics and Probability Letters, **119** (2016), 344–348. MR3555307
[11] M. Chebunin, A. Kovalevskii, *Asymptotically normal estimators for Zipf's law*, Sankhya A (2018).
[12] G. Decrouez, M. Grabchak, Q. Paris, *Finite sample properties of the mean occupancy counts and probabilities*, Bernoulli **24** (2018), no. 3, 1910–1941 MR3757518
[13] P. Deheuvels, G.V. Martynov, *Cramer-von mises-type tests with applications to tests of independence for multivariate extreme-value distributions*, Communications in Statistics — Theory and Methods, **25**:4 (1996), 871–908. MR1380624
[14] O. Durieu, Y. Wang, *From infinite urn schemes to decompositions of self-similar Gaussian processes*, Electron. J. Probab. **21** (2016), paper no. 43, 23 pp. MR3530320
[15] O. Durieu, G. Samorodnitsky, Y. Wang, *From infinite urn schemes to self-similar stable processes*, Stochastic Processes and their Applications (2019, in press).
[16] M. Dutko, *Central limit theorems for infinite urn models*, Ann. Probab., **17** (1989), 1255–1263. MR1009456
[17] M. Gerlach, E.G. Altmann, *Stochastic Model for the Vocabulary Growth in Natural Languages*, Physical Review X, **3**, 021006 (2013).
[18] A. Gnedin, B. Hansen, J. Pitman, *Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws*, Probability Surveys, **4** (2007), 146–171. MR2318403
[19] A. Guillou, P. Hall, *A diagnostic for selecting the threshold in extreme value analysis*, Journal of the Royal Statistical Society: Series B., **63**:2, 293–305 (2002). MR1841416
[20] H.S. Heaps, *Information Retrieval: Computational and Theoretical Aspects*, Academic Press, 1978.
[21] G. Herdan, *Type-token mathematics*, The Hague: Mouton, 1960.
[22] H.-K. Hwang, S. Janson, *Local Limit Theorems for Finite and Infinite Urn Models*, The Annals of Probability **36**:3 (2008), 992–1022. MR1620350
[23] I. Eliazar. *The Growth Statistics of Zipfian Ensembles: Beyond Heaps' Law*, Physica (Amsterdam) **390**, 3189 (2011).
[24] S. Karlin, *Central Limit Theorems for Certain Infinite Urn Schemes*, Jounal of Mathematics and Mechanics, **17**:4 (1967), 373–401. MR0216548
[25] E. S. Key, *Rare Numbers*, Journal of Theoretical Probability **5**:2 (1992), 375–389. MR1157991
[26] E. S. Key, *Divergence rates for the number of rare numbers*, Journal of Theoretical Probability **9**:2 (1996), 413–428. MR1385405
[27] A.P. Kovalevskii, E.V. Shatalin, *Asymptotics of sums of residuals of one-parameter linear regression on order statistics*, Theory of probability and its applications, **59**:3 (2015), 375–387. MR3415974
[28] A. Kovalevskii, E. Shatalin, *A limit process for a sequence of partial sums of residuals of a simple regression on order statistics* , Probability and Mathematical Statistics, **36**, Fasc. 1 (2016), 113–120. MR3529343
[29] D.C. van Leijenhorst, T.P. van der Weide, *A Formal Derivation of Heaps' Law*, Information Sciences (NY) **170**, 263 (2005).
[30] G.V. Martynov, *Omega-square tests*, Nauka, Moscow, 1978 (in Russian). MR0527912
[31] A. Muratov, S. Zuyev, *Bit flipping and time to recover*, J. Appl. Probab. **53**:3 (2016), 650–666. MR3570086
[32] P.T. Nicholls, *Estimation of Zipf parameters*, J. Am. Soc. Inf. Sci., **38** (1987), 443–445.
[33] M.I. Ohannessian, M.A. Dahleh, *Rare probability estimation under regularly varying heavy tails*, Proceedings of the 25th Annual Conference on Learning Theory, PMLR 23:21.1–21.24 (2012).
[34] A.M. Petersen, J.N. Tenenbaum, S. Havlin, H.E. Stanley, M. Perc, *Languages cool as they expand: Allometric scaling and the decreasing need for new words*, Scientific Reports **2**, Article No. 943 (2012).
[35] M.A. Serrano, A. Flammini, F. Menczer, *Modeling Statistical Properties of Written Text*, PLoS ONE **4**, e5372 (2009).

[36] N.V. Smirnov, *On the omega-squared distribution*, Mat. Sb. **2**, 973–993 (1937, in Russian).
[37] N.S. Zakrevskaya, A.P. Kovalevskii, *One-parameter probabilistic models of text statistics*, Sib. Zh. Ind. Mat., **4**:2 (2001), 142–153 (in Russian). MR1965927
[38] N. Zakrevskaya, A. Kovalevskii, *An omega-square statistics for analysis of correspondence of small texts to the Zipf—Mandelbrot law*, Applied methods of statistical analysis. Statistical computation and simulation — AMSA'2019, 18–20 September 2019, Novosibirsk: Proceedings of the International Workshop, Novosibirsk: NSTU (2019), 488–494.
[39] G.K. Zipf, *The Psycho-Biology of Language*, Routledge, London, 1936.

Mikhail Georgievich Chebunin
Sobolev Institute of Mathematics,
4, Koptyuga ave.,
Novosibirsk State University,
1, Pirogova str.,
Novosibirsk, 630090, Russia
E-mail address: chebuninmikhail@gmail.com

Artyom Pavlovich Kovalevskii
Novosibirsk State Technical University,
20, K. Marksa ave.,
630073, Novosibirsk, Russia
Novosibirsk State University,
1, Pirogova str.,
Novosibirsk, 630090, Russia
E-mail address: artyom.kovalevskii@gmail.com