Hapax legomena via stochastic processes

Shahzod Fayzullaev[1,2] (0009-0005-5402-2306), Artyom Kovalevskii[1,3,4*] (0000-0001-5808-3134)

[1] Novosibirsk State University, Novosibirsk, Russia
[2] Urgench State University, Urgench, Uzbekistan
[3] Sobolev Institute of Mathematics, Novosibirsk, Russia
[4] Novosibirsk State Technical University, Novosibirsk, Russia
[*] Corresponding author's email: artyom.kovalevskii@gmail.com

DOI: https://doi.org/10.53482/2024_56_415

ABSTRACT

We study the number of words that occur exactly once since the beginning of a text. We model it as a stochastic process over the length of the text. The elementary probability model, going back to Bahadur and Karlin, states that the number of words that occur exactly once should grow according to a power law, like the number of different words. The final value of the number of words occurring exactly once is the number of hapaxes of this text. We construct two statistical tests to test Karlin's model under the assumption that the probabilities of words in this model satisfy the generalized Zipf's law. These statistical tests show that some texts fit the model well, but many texts deviate significantly from it. This deviation is that the number of hapaxes is too small relative to the number of different words.

Keywords: Zipf's law, statistical test, mathematical expectation, limit theorems.

1 Introduction

Hapax legomena are words that appear in a text (or a corpus of texts) only once. Harrison (1921) noted that the number of hapax legomena in a text characterises an author's style. But the proper dependence of the number of hapax legomena from the number of all words in a text was unclear.

Karlin (1967) studied a general infinite box scheme: each ball, independently of others, chooses one of infinitely many boxes with correspondence to some probability distribution (the same for all balls).

This infinite box scheme can be interpreted as an elementary probabilistic text model: boxes correspond to words of an infinite dictionary, balls correspond to sequential words of a text. So words are chosen randomly and independently in this model.

Karlin studied a very general case of the probability distribution that generalise the Zipf's Law.

Among others, he studied statistics R_n of the number of non-empty urns after n balls. This statistics corresponds to the number of different words in the elementary probabilistic text model for a text of n words.

Another statistics are $R_{n,1}$, $R_{n,2}$, ..., of numbers of urns with exactly one ball, two balls, etc. These statistics are interpreted as the numbers of words that present only once in the text (hapax legomena of this text), only twice (dis legomena), etc.

Popescu and Altmann (2008) studied language typology on the base of the number of different words R_n of a text, that have been designated as V, vocabulary, and the number of words that appear only once $R_{n,1}$ in a text, that have been designated as HL, hapax legomena. They found that $R_{n,1}/R_n$ is almost constant for a collection of texts in different languages.

Research on the correspondence of texts to the Zipf's law has not stopped for more than a hundred years. Here are some principal and recent studies. Balasubrahmanyan and Naranan (1996) studied the Zipf's law using a unified entropy model for two classes of words: content words and service words. There are many another explanations of Zipf's law, see Manin (2009) and cited papers therein.

van Egmond et al. (2015) established the correspondence of the speech of people with non-fluent aphasia to Zipf's law. They show no worse fit for aphasic compared to healthy speech but only a difference in slope.

Tuzzi et al. (2009) performed a very interesting analysis of a corpus composed of the End of Year addresses by Italian presidents. The results show that Zipf's law is an appropriate model. This is interesting because sometimes the texts were compiled by two persons or more. We see, one can apply the analysis of these texts as processes, this can give non-trivial results.

Complex lexicons that combine lexicons of different authors do not correspond to the Zipf's law. Cancho and Solé (2001) have studied complex lexicons and found that the classical Zipf's law cannot be applied for its approximation, there are two diapasons with different Zipf exponents.

A recent study of Grabchak et al. (2020) deals with a model with switching of regimes. This is a variant of a model for complex lexicons.

A lexicon of any one author corresponds to the Zipf's law, and many of the correspondences can be found in the literature.

Bahadur (1960) was the first who explained the law of increasing of number of different words on the base of the Zipf's law and an elementary probability model of an infinite urn scheme.

We use the next modification of the model by Karlin (1967). It is known as Karlin balls-in-boxes model, and we refer to it as the elementary text model.

Independent identically distributed random variables $X_1, X_2, ..., X_n$ have an integer-valued power distri-

bution:

$$p_k := \mathbf{P}(X_1 = k) = l(\alpha, k)k^{-\alpha}, \quad k \geq 1, \tag{1}$$

where $\alpha > 1$ is an unknown parameter, $l(\alpha, k) > 0$ is a function such that the normalization condition

$$\sum_{k=1}^{\infty} l(\alpha, k)k^{-\alpha} = 1,$$

holds and $l(\alpha, x)$ is a slowly varying function as $x \to \infty$ for any $\alpha > 1$, that is, $l(\alpha, cx)/l(\alpha, x) \to 1$ for any $c > 0$.

We assume $p_1 \geq p_2 \geq \ldots$.

α is the Zipf exponent.

Examples of such distributions are:

1. $p_k = k^{-\alpha}/\zeta(\alpha)$, where $\zeta(\alpha) = \sum_{k=1}^{\infty} k^{-\alpha}$ be the Riemann zeta function (this is an infinite Zipf model);
2. $p_k = k^{1-\alpha} - (k+1)^{1-\alpha}$ (this is a difference Zipf model);
3. $p_k = (k+q)^{-\alpha}/\zeta(\alpha, q+1)$, where $\zeta(\alpha, q+1) = \sum_{k=0}^{\infty}(k+q+1)^{-\alpha}$ be the Hurvitz zeta function (this is an infinite Zipf—Mandelbrot model), $q > -1$ be the Mandelbrot shift;
4. $p_k = (1+q)^{\alpha-1}\left((k+q)^{1-\alpha} - (k+q+1)^{1-\alpha}\right)$ (this is a difference Zipf-Mandelbrot model).

Random variables $X_1, \ldots, X_n$ correspond to sequential words of a text, and their values $1, 2, \ldots$ correspond to words of an infinite vocabulary.

We consider the number R_n of different words and the number $R_{n,j}$ of words that occurred exactly j times in a text of n words.

Statistics $R_{n,1}$ is the number of words that occurred only once (hapax legomena of the text).

Note that $R_n = \sum_{j=1}^{n} R_{n,j}$.

Karlin (1967) proved, in particular, the strong law of large numbers and the central limit theorem for R_n and $R_{n,1}$.

Key (1992) studied the asymptotics of the number of unique words. In particular, he showed that if $\lim_{i\to\infty} p_{i+1}/p_i = 1$ then $R_{n,1} \to \infty$ in probability as $n \to \infty$. If $\limsup_{i\to\infty} p_{i+1}/p_i < 1$ then $\mathbf{E}R_{n,1}$ is uniformly bounded.

The statistics R_n and $R_{n,1}$ in these models are also studied in the works of Bahadur (1960), Barbour (2009), Barbour and Gnedin (2009), Ben-Hamou et al. (2017), Chebunin and Kovalevskii (2016), Chebunin and

Kovalevskii (2019), Decrouez et al. (2018), Dutko (1989), Gnedin et al. (2007), Gnedin et al. (2010), Zakrevskaya and Kovalevskii (2001).

Let illustrate the process of numbers of different words and the process of numbers of words that counted only once. We drew them for the novel *Peter Pan* by J. M. Barrie (https://gutenberg.org/ebooks/16), see Fig. 1.

The beginning of the text is:

'Chapter I. PETER BREAKS THROUGH

All children, except one, grow up. They soon know that they will grow up, and the way Wendy knew was this.'

We converted the text to lowercase, lemmatized it and removed all punctuation marks and non-letter characters:

1:'chapter', 2:'i', 3:'peter', 4:'break', 5:'through', 6:'all', 7:'child', 8:'except', 9:'one', 10:'grow', 11:'up', 12:'they', 13:'soon', 14:'know', 15:'that', 16:'they', 17:'will', 18:'grow', 19:'up', 20:'and', 21:'the', 22:'way', 23:'wendy', 24:'know', 25:'be', 26:'this'

We calculate numbers of different words and numbers of words that counted once from the beginning of the text:

	chapter	i	...	**grow**	**up**	**they**	soon	**know**	that	**they**	will	**grow**	**up**	and	the	way	wendy	**know**	be	this
k	1	2	...	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
R_k	1	2	...	10	11	12	13	14	15	15	16	16	16	17	18	19	20	20	21	22
$R_{k,1}$	1	2	...	10	11	12	13	14	15	14	15	14	13	14	15	16	17	16	17	18

Word 16:'they' coincides with word 12:'they', therefore the number of different words does not ascend, $R_{16} = R_{15}$, and the number of words that counted only once descends, $R_{16,1} = R_{15,1} - 1$. The same is true for words 18:'grow', 19:'up', 24:'know', that coincide with words 10:'grow', 11:'up', 14:'know'.

The whole text consists of $n = 47987$ words, half the number of words of the text is $\lfloor n/2 \rfloor = 23993$. The number of different words in the whole text is $R_n = 3719$, the number of different words in the first half of the text is $R_{\lfloor n/2 \rfloor} = 2487$. The number of words that counted only once in the whole text (that is, the number of hapax legomena of the text) is $R_{n,1} = 1705$, the number of words that counted only once in the first half of the text is $R_{\lfloor n/2 \rfloor,1} = 1187$, see Fig. 1.

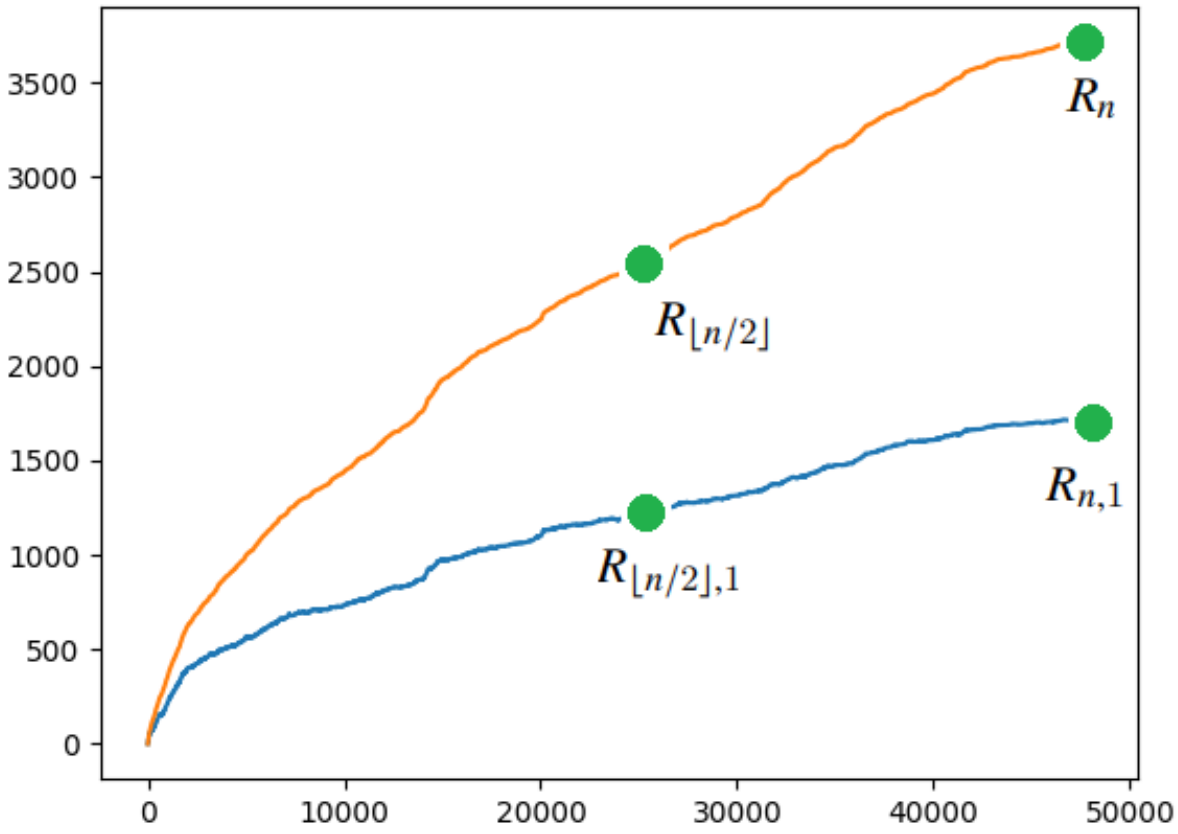

Figure 1: A diagram for *Peter Pan*

Hapax legomena of this text include short words as 'sly', 'pat', 'wag' and longer ones as 'cauliflower', 'vindictiveness', 'beseechingly'.

Numbers of different words and numbers of words encounted only once follow similar power laws.

So $R_{n,1}/R_n$ and $R_{\lfloor n/2 \rfloor,1}/R_{\lfloor n/2 \rfloor}$ are almost the same, this gives the idea of statistics in our Theorem 1.

On the other hand, $R_{\lfloor n/2 \rfloor}/R_n$ and $R_{\lfloor n/2 \rfloor,1}/R_{n,1}$ are almost the same, this gives the idea of statistics in our Theorem 2.

2 Material

French and German texts were taken from the most popular of all time on https://gutenberg.org/, while English texts were taken from the most popular of the last 30 days on https://gutenberg.org/. Russian texts were downloaded from https://all-the-books.ru/, this list of texts was compiled based on the article "What is the most popular work of a Russian writer in the world?" on https://www.cultural.ru/ and the best books section on https://www.livelib.ru/. Lemmatization of words in texts was carried out using the

"spacy" library of the Python programming language.

3 Methodology

Our theoretical analysis is based on the elementary text model with assumption (1).

We designate $\theta = 1/\alpha$. So θ is the inverse Zipf exponent, $0 < \theta < 1$.

Formula (1) expresses the generalized power law $p_k = l(\alpha, k)k^{-\alpha}$. It is equivalent to (see Karlin (1967))

$$\kappa(x) := \max(k > 0 : \ p_k \geq 1/x) = L(\theta, x)x^{\theta} \tag{2}$$

with a slowly varying function $L(\theta, x)$, that is, for any $0 < \theta < 1$, $c > 0$ we have $L(\theta, cx)/L(\theta, x) \to 1$ as $x \to \infty$.

Karlin proved convergences $R_n/\mathbf{E}R_n \to 1$, $R_{n,1}/\mathbf{E}R_{n,1} \to 1$ a.s. (almost surely). We will refer to this fact as the Strong Law of Large Numbers (SLLN) for R_n and $R_{n,1}$.

Karlin found asymptotics

$$\mathbf{E}R_n \sim \Gamma(1-\theta)L(\theta, n)n^{\theta}, \quad \mathbf{E}R_{n,1} \sim \theta\Gamma(1-\theta)L(\theta, n)n^{\theta} \tag{3}$$

as $n \to \infty$.

Chebunin and Kovalevskii (2019) proposed estimate

$$\widehat{\theta} = R_{n,1}/R_n, \tag{4}$$

that is, a fraction of the hapax legomena in all the different words of a text. From SLLN and (3), this estimate is strongly consistent, $\widehat{\theta} \to \theta$ a.s. as $n \to \infty$.

So our idea is to compare numbers of different words and words that encounted only once in the whole text, R_n and $R_{n,1}$, with the same quantities in the first half of the text, $R_{\lfloor n/2 \rfloor}$ and $R_{\lfloor n/2 \rfloor,1}$, see Fig. 1.

We use the Functional Central Limit Theorem for the vector process $(R_{\lfloor nt \rfloor}, \ R_{\lfloor nt \rfloor,1})$, $\quad 0 \leq t \leq 1$.

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

$$Y_n(t) = \frac{R_{[nt]} - \mathbf{E}R_{[nt]}}{\sqrt{\mathbf{E}R_n}}, \qquad Y_{n,k}(t) = \frac{R_{[nt],k} - \mathbf{E}R_{[nt],k}}{\sqrt{\mathbf{E}R_n}}.$$

The following theorem was proved by Chebunin and Kovalevskii (2016):

If (1) is satisfied, $\theta = 1/\alpha$, then for any integer $\nu \geq 1$ the random process $\{(Y_n(t), Y_{n,1}(t), ..., Y_{n,\nu}(t)), \ 0 \leq t \leq 1\}$ converges weakly in the uniform metric in $D(0, 1)$ to $(\nu + 1)$ -dimensional Gaussian process

$$\{(Y(t), Y_1(t), ..., Y_{\ \nu}(t)), \ 0 \leq t \leq 1\}$$

with continuous trajectories, zero expectation and covariance function $(c_{ij}(\tau,t))_{i,j=0}^{\nu}$,

$c_{ij}(\tau,t) = \frac{\theta\tau^i(t-\tau)^{j-i}t^{\theta-j}\Gamma(j-\theta)}{i!(j-i)!\Gamma(1-\theta)} - \frac{\theta\tau^i t^j(t+\tau)^{\theta-i-j}\Gamma(i+j-\theta)}{i!j!\Gamma(1-\theta)}$ for $1 \le i < j, \tau \le t$,

$c_{ij}(\tau,t) = -\frac{\theta\tau^i t^j(t+\tau)^{\theta-i-j}\Gamma(i+j-\theta)}{i!j!\Gamma(1-\theta)}$ for $i > j \ge 1, \tau \le t$,

$c_{ii}(\tau,t) = \frac{\theta\tau^\theta\Gamma(i-\theta)}{i!\Gamma(1-\theta)} - \frac{\theta\tau^i t^i(t+\tau)^{\theta-2i}\Gamma(2i-\theta)}{(i!)^2\Gamma(1-\theta)}$ for $i > 0, \tau \le t$,

$c_{00}(\tau,t) = \left((t+\tau)^\theta - t^\theta\right)$ for $\tau \le t$,

$c_{i0}(\tau,t) = -\frac{\theta\tau^i(t+\tau)^{\theta-i}\Gamma(i-\theta)}{i!\Gamma(1-\theta)}$ for $i > 0, \tau \le t$,

$c_{0j}(\tau,t) = \frac{\theta((t-\tau)^j t^{\theta-j} - t^j(t+\tau)^{\theta-j})\Gamma(j-\theta)}{j!\Gamma(1-\theta)}$ for $j > 0, \tau \le t$,

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

The next lemma is a straightforward consequence of the theorem. This lemma is a base of our theorems that give a new method of statistical analysis of texts.

Lemma 1 If (1) is satisfied, $\theta = 1/\alpha$, then the random vector $\left(Y_n(1), Y_n(1/2), Y_{n,1}(1), Y_{n,1}(1/2)\right)$ converges weakly to the normal vector $(Y(1), Y(1/2), Y_1(1), Y_1(1/2))$ with zero expectation and covariance matrix $G(\theta) = (g_{ij}(\theta))_{i,j=1}^4$, $g_{ji}(\theta) = g_{ij}(\theta)$,

$$g_{11}(\theta) = c_{00}(1,1) = 2^\theta - 1,$$

$$g_{12}(\theta) = c_{00}(1/2,1) = (3/2)^\theta - 1,$$

$$g_{13}(\theta) = c_{01}(1,1) = -\theta 2^{\theta-1},$$

$$g_{14}(\theta) = c_{10}(1/2,1) = -\theta\left(1/2 - (3/2)^{\theta-1}\right),$$

$$g_{22}(\theta) = c_{00}(1/2,1/2) = 1 - (1/2)^\theta,$$

$$g_{23}(\theta) = c_{01}(1/2,1) = \theta\left(1/2 - (3/2)^{\theta-1}\right),$$

$$g_{24}(\theta) = c_{01}(1/2,1/2) = -\theta/2,$$

$$g_{33}(\theta) = c_{11}(1,1) = \theta\left(1 - 2^{\theta-2}(1-\theta)\right),$$

$$g_{34}(\theta) = c_{11}(1/2,1) = \theta(1/2)^\theta\left(1 - (3/2)^{\theta-2}(1-\theta)\right),$$

$$g_{44}(\theta) = c_{11}(1/2,1/2) = \theta\left((1/2)^\theta - (1-\theta)/4\right).$$

So we test the elementary probability text model comparing fractions of words that occur only once in the whole text and in the first half of the text. We use Lemma 1 to calculate the limiting distribution. We need a more restrictive condition on probabilities:

$$p_k = ck^{-1/\theta}(1 + o(k^{-1/2})) \text{ as } k \to \infty,\ \ c > 0,\ 0 < \theta < 1. \tag{5}$$

Note that examples in Introduction (infinite Zipf, Zipf-Mandelbrot models and their modifications) satisfy this condition.

The idea of Theorem 1 is to compare the fraction of hapaxes in all different words for the whole text and for the first half of the text. The limiting normal distribution has variance that is calculated using matrix $G(\theta)$ from Lemma 1.

Theorem 1 If (5) holds then the statistics

$$V_n^{(1)} = \sqrt{R_n}\left(\frac{R_{n,1}}{R_n} - \frac{R_{\lfloor n/2\rfloor,1}}{R_{\lfloor n/2\rfloor}}\right)$$

converges weakly to a normal random variable $V^{(1)} = Y_1(1) - \theta Y(1) - 2^\theta Y_1(1/2) + \theta 2^\theta Y(1/2)$ with zero expectation and variance $b_1^2(\theta) = m_1(\theta)G(\theta)m_1^T(\theta)$, where $m_1(\theta) = (-\theta, \theta 2^\theta, 1, -2^\theta)$.

The idea of Theorem 2 is to compare the fraction of different words in the first half of the text in different words of the whole text with the analogous fraction for words encounted only once.

Theorem 2 If (5) holds then the statistics

$$V_n^{(2)} = \sqrt{R_n}\left(\frac{R_{\lfloor n/2\rfloor}}{R_n} - \frac{R_{\lfloor n/2\rfloor,1}}{R_{n,1}}\right)$$

converges weakly to a normal random variable $V^{(2)} = Y(1/2) - 2^{-\theta}Y(1) - \theta^{-1}Y_1(1/2) + 2^{-\theta}Y_1(1)$ with zero expectation and variance $b_2^2(\theta) = m_2(\theta)G(\theta)m_2^T(\theta)$, where $m_2(\theta) = (-2^{-\theta}, 1, 2^{-\theta}, -\theta^{-1})$.

Proofs of Theorems 1 and 2 are in Appendix.

Based on Theorems 1 and 2, statistical tests of the elementary probability model are constructed respectively as follows:

$$\text{p-value}^{(1)} = 2\Phi^{-1}\left(-\left|\frac{\sqrt{R_n}}{b_1(\widehat{\theta})}\left(\frac{R_{n,1}}{R_n} - \frac{R_{\lfloor n/2\rfloor,1}}{R_{\lfloor n/2\rfloor}}\right)\right|\right),$$

$$\text{p-value}^{(2)} = 2\Phi^{-1}\left(-\left|\frac{\sqrt{R_n}}{b_2(\widehat{\theta})}\left(\frac{R_{\lfloor n/2\rfloor}}{R_n} - \frac{R_{\lfloor n/2\rfloor,1}}{R_{n,1}}\right)\right|\right),$$

where Φ^{-1} is the quantile function of the standard normal distribution, $\widehat{\theta}$ is given by (4), $b_n^{(j)} = b_j(\widehat{\theta})$, $j = 1, 2$.

Let us denote

$$Q_n^{(1)} = \frac{\sqrt{R_n}}{b_1(\widehat{\theta})}\left(\frac{R_{n,1}}{R_n} - \frac{R_{\lfloor n/2\rfloor,1}}{R_{\lfloor n/2\rfloor}}\right),$$

$$Q_n^{(2)} = \frac{\sqrt{R_n}}{b_2(\widehat{\theta})}\left(\frac{R_{\lfloor n/2\rfloor}}{R_n} - \frac{R_{\lfloor n/2\rfloor,1}}{R_{n,1}}\right).$$

If $\text{p-value}^{(j)} < \varepsilon$ then the hypothesis of the elementary probability model is rejected at level ε.

If p-value$^{(j)} \geq \varepsilon$ then this hypothesis is accepted at level of ε. Here j=1, 2 be the number of the test corresponding to Theorem 1, 2.

4 Application to text analysis

Analysis of novels in English, French, German and Russian is presented in Table 1 and Table 2. So, Table 1 contains text statistics and estimate $\widehat{\theta}$, and Table 2 contains $Q_n^{(1)}$ and $Q_n^{(1)}$ and calculations of p-values on the base of Theorems 1 and 2.

Authors and translators (if the original text is in a foreign language) of novels can be found using the appropriate links in the "Material" section.

There is dependence of estimate $\widehat{\theta}$ on the language. Remember that the estimate of the Hurst exponent is the inverse value, $\widehat{\alpha} = 1/\widehat{\theta} = R_n/R_{n,1}$. Maximal $\widehat{\theta}$ (and minimal $\widehat{\alpha}$ are for texts in German, lesser for Russian, even lesser for English, and the minimal $\widehat{\theta}$ and maximal estimates of Hurst exponents are for French texts, see Popescu and Altmann (2008) for the comparison with another Hurst exponent estimates.

The analysis of Table 2 shows that the first fraction in both statistics $Q_n^{(1)}$ and $Q_n^{(2)}$ is typically greater than the second one, so the statistics are negative.

The corresponding p-values depend on the absolute values of the statistics. The p-values are small in many cases, and this means bad correspondence of real texts and the elementary text model. There are many almost zero p-values for English and French texts. All the p-values for French texts are lesser than 0.1. The p-values are not so small for German and Russian texts. But the accurate correspondence to the probability model assumes that about 90% of p-values are greater than 0.1, but this is not hold for any language under consideration.

This bad correspondence of real texts to the elementary probability text model (known as the Karlin occupancy scheme or Karlin model) means that this model needs modifications. Some variants of modifications have been studied by Chebunin and Kovalevskii (2022).

A typical deviation of real texts from the elementary probabilistic model is as follows. The appearance of new words throughout the text follows a power law, but the number of words that appear exactly once grows much more slowly. Therefore, a significant discrepancy arises in the increase in the number of different words and the number of words that appeared only once.

An example of this deviation from the model is *The Scarlet Letter* by Nathaniel Hawthorne, see Figure 2. Here the number of different words corresponds to the power law; it increases from 5098 for half the text to 6571 for the entire text, that is, by 29%. The number of words that appeared only once is 2316 for half the text and 2630 for the entire text, that is, it increases by 14%. This difference does not correspond

Table 1: Novels and their statistics

Novel	n	R_n	$R_{\lfloor n/2 \rfloor}$	$R_{n,1}$	$R_{\lfloor n/2 \rfloor,1}$	$\widehat{\theta}$
Alice's Adventures in Wonderland	28033	2000	1455	758	635	0.3790
Dracula	166346	7252	5342	3118	2462	0.4300
Frankenstein; Or, The Modern Prometheus	75567	5295	4037	2038	1758	0.3849
Metamorphosis	22387	2049	1433	952	681	0.4646
Pride and Prejudice	129690	5193	4067	1917	1657	0.3692
Romeo and Juliet	26837	3082	2099	1587	1130	0.5149
Simple Sabotage Field Manual	9232	1838	1162	918	631	0.4995
The Great Gatsby	50168	4764	3505	2362	1925	0.4958
The Picture of Dorian Gray	80571	5306	3430	2333	1540	0.4397
The Scarlet Letter	86018	6571	5098	2630	2316	0.4002
Arsène Lupin, gentleman-cambrioleur	59617	5024	3727	2105	1781	0.4190
Aventures d'Alice au pays des merveilles	28725	2317	1702	1009	833	0.4355
Contes Français	81547	6631	4521	3062	2240	0.4618
Discours de la méthode	126042	4839	4147	1940	1744	0.4009
Germinal	177163	7189	5469	2415	2088	0.3359
L'Assommoir	171698	6969	5156	2547	2024	0.3655
Le blé en herbe: roman	32627	4013	2790	1997	1530	0.4976
Le Fantôme de l'Opéra	109602	5559	4224	2222	1820	0.3997
Le Grand Meaulnes	68831	4577	3292	1902	1506	0.4156
Le Horla	44535	4027	2762	1791	1349	0.4447
Madame Bovary	120404	7825	5643	3270	2593	0.4179
Monsieur Vénus	44524	4733	3450	2230	1803	0.4712
Alice's Abenteuer im Wunderland	25485	3137	2108	1702	1204	0.5426
Also sprach Zarathustra	84708	7691	4750	4207	2613	0.5470
Anna Karenina	189690	13489	8957	7180	4914	0.5323
Buddenbrooks	230238	20129	12759	10826	7235	0.5378
Demian	47216	5499	3494	3204	2075	0.5827
Der Zauberberg. Erster Band	147383	17240	10022	10278	6044	0.5962
Der Zauberberg. Zweiter Band	159852	19608	12317	11860	7765	0.6049
Faust. Erster Teil	31048	5234	3410	3249	2201	0.6207
In Stahlgewittern	67740	10426	6843	6320	4311	0.6062
Unterm Rad	49258	7897	5056	4943	3296	0.6259
Anna Karenina	286673	18774	12174	9323	6181	0.4966
Bratya Karamazovy	297489	19169	13416	9335	6851	0.4870
Doktor Zhivago	156761	22689	14635	12918	8706	0.5694
Geroy nashego vremeni	42537	6452	4414	3363	2465	0.5212
Idiot	211861	13983	9797	6753	4956	0.4829
Mertvye dushi	113147	14312	9050	8049	5307	0.5624
Odin den Ivana Denisovicha	32852	6224	4004	3775	2519	0.6065
Otcy i deti	55297	8170	5381	4477	3042	0.5480
Prestuplenie i nakazanie	197679	16658	10366	8234	5421	0.4943
Revizor	21600	3593	2372	1982	1352	0.5516

to the elementary model of the text, according to which these percentages should be almost equal. The appearance of new words in the second half of the text is largely compensated by the fact that words appear a second time that appeared only once in the first half of the text.

An interesting example of a non-homogenious text is *The Picture of Dorian Gray* by Oscar Wilde, see Figure 3.

Table 2: Test statistics and p-values for novels

Novel	$b_n^{(1)}(\widehat{\theta})$	$Q_n^{(1)}$	p-value$^{(1)}$	$b_n^{(2)}(\widehat{\theta})$	$Q_n^{(2)}$	p-value$^{(2)}$
Alice's Adventures in Wonderland	0.5608	-4.5791	0.0000	1.6168	-3.0490	0.0023
Dracula	0.6532	-4.0319	0.0001	1.5335	-2.9423	0.0033
Frankenstein; Or, The Modern Prometheus	0.5713	-6.4420	0.0000	1.6064	-4.5385	0.0000
Metamorphosis	0.7182	-0.6688	0.5037	1.4838	-0.4872	0.6261
Pride and Prejudice	0.5434	-5.0756	0.0000	1.6347	-3.5797	0.0003
Romeo and Juliet	0.8157	-1.5944	0.1108	1.4193	-1.2119	0.2255
Simple Sabotage Field Manual	0.7853	-2.3789	0.0174	1.4383	-1.644	0.1002
The Great Gatsby	0.7781	-4.7378	0.0000	1.4429	-3.7914	0.0001
The Picture of Dorian Gray	0.6713	-1.008	0.3135	1.519	-0.6549	0.5126
The Scarlet Letter	0.5989	-7.316	0.0000	1.5804	-5.3742	0.0000
Arsène Lupin, gentleman-cambrioleur	0.633	-6.5925	0.0000	1.5503	-4.766	0.0000
Aventures d'Alice au pays des merveilles	0.6634	-3.9142	0.0001	1.5252	-2.8719	0.0041
Contes Français	0.7128	-3.8496	0.0001	1.4877	-2.7231	0.0065
Discours de la méthode	0.6001	-2.2761	0.0228	1.5793	-1.8489	0.0645
Germinal	0.4856	-8.0069	0.0000	1.7001	-5.1793	0.000
L'Assommoir	0.537	-4.2097	0.0000	1.6415	-2.7876	0.0053
Le blé en herbe: roman	0.7817	-4.1130	0.0000	1.4406	-3.1181	0.0018
Le Fantôme de l'Opéra	0.5979	-3.8853	0.0001	1.5812	-2.7930	0.0052
Le Grand Meaulnes	0.6267	-4.5248	0.0000	1.5557	-3.1551	0.0016
Le Horla	0.6807	-4.0708	0.0000	1.5117	-2.8268	0.0047
Madame Bovary	0.6310	-5.8341	0.0000	1.5520	-4.0933	0.0000
Monsieur Vénus	0.7306	-4.8445	0.0000	1.4750	-3.7126	0.0002
Alice's Abenteuer im Wunderland	0.8709	-1.8393	0.0659	1.387	-1.4304	0.1526
Also sprach Zarathustra	0.8800	-0.3092	0.7572	1.3820	-0.2223	0.8241
Anna Karenina	0.8503	-2.2314	0.0257	1.3988	-1.692	0.0906
Buddenbrooks	0.8614	-4.8126	0.0000	1.3924	-3.5089	0.0004
Demian	0.9534	-0.873	0.3827	1.3432	-0.6757	0.4992
Der Zauberberg. Erster Band	0.9818	-0.9230	0.3560	1.3292	-0.6648	0.5062
Der Zauberberg. Zweiter Band	1.0002	-3.5805	0.0003	1.3203	-2.8169	0.0048
Faust. Erster Teil	1.0342	-1.7282	0.0839	1.3045	-1.4381	0.1504
In Stahlgewittern	1.0030	-2.4240	0.0154	1.3190	-1.9958	0.0460
Unterm Rad	1.0454	-2.2072	0.0273	1.2994	-1.8164	0.0693
Anna Karenina	0.7797	-1.9560	0.0505	1.4419	-1.3811	0.1673
Bratya Karamazovy	0.761	-4.3071	0.0000	1.4542	-3.2395	0.0012
Doktor Zhivago	0.9257	-4.1532	0.0000	1.3574	-3.209	0.0013
Geroy nashego vremeni	0.8282	-3.6095	0.0003	1.4118	-2.7792	0.0054
Idiot	0.7532	-3.5991	0.0003	1.4594	-2.6949	0.0070
Mertvye dushi	0.9114	-3.1521	0.0016	1.3649	-2.3665	0.0180
Odin den Ivana Denisovicha	1.0037	-1.7762	0.0757	1.3186	-1.4340	0.1516
Otcy i deti	0.8819	-1.7773	0.0755	1.3809	-1.3644	0.1724
Prestuplenie i nakazanie	0.7752	-4.772	0.0000	1.4448	-3.2234	0.0013
Revizor	0.8894	-1.2371	0.2160	1.3768	-0.9564	0.3389

Plots of the number of different words and the number of words occurring exactly once behave roughly like power functions up to 50,000 words. At this moment, many new words begin to appear in the text. On the graph, this looks like a wave, since the growth rate of the number of different words and words occurring exactly once increases. This wave is similar to the one at the beginning of the text. This behavior of the graph indicates the presence of a change point, that is, a violation of the probabilistic properties of the text. The presence of a change point can be diagnosed by analyzing graphs in the forward and backward direction of the text, as in Abebe et al. (2022). The position of the change point can be determined by the method described in Abebe et al. (2023). This position roughly corresponds to

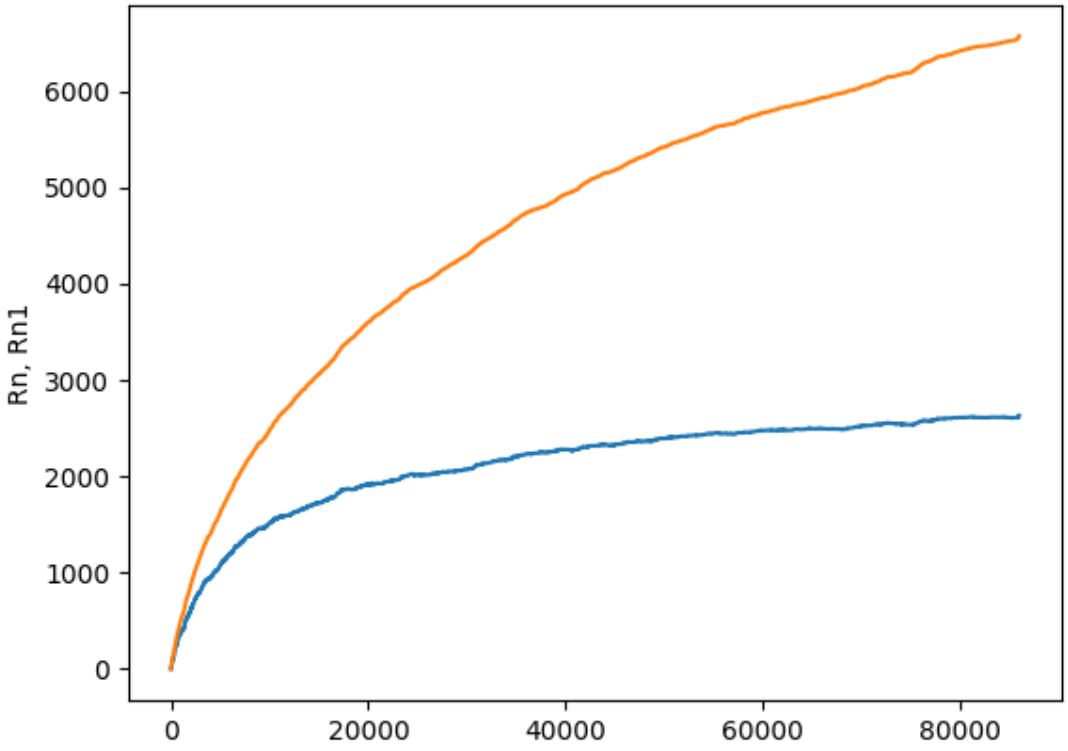

Figure 2: *The Scarlet Letter* by Nathaniel Hawthorne

the beginning of Chapter 11. Please note that there is almost no dialogue in Chapter 11. Apparently, it is the lack of dialogue that leads to this relatively rapid increase. The author's text, compared to dialogues, contains a significantly smaller number of stereotypically repeated words.

P-values depend significantly on the length of the text. A general fact of mathematical statistics is that the smaller the difference between the actual distribution and the theoretical distribution, the longer the sequence length required to detect this difference. Thus, among the 42 texts studied (Tables 1 and 2), the p-values calculated using Theorem 2 are greater than 0.05 for 18 texts. If we consider 11 texts longer than 150,000 words, we see that only one of them (*Anna Karenina* by Leo Tolstoy in Russian) reaches a p-value greater than 0.05. All 5 of the shortest texts studied (less than 27,000 words in length) have p-values greater than 0.1. Thus, the differences between real texts and an elementary probabilistic model are, as a rule, insignificant for short texts (stories, chapters of novels), but they are significant for long ones.

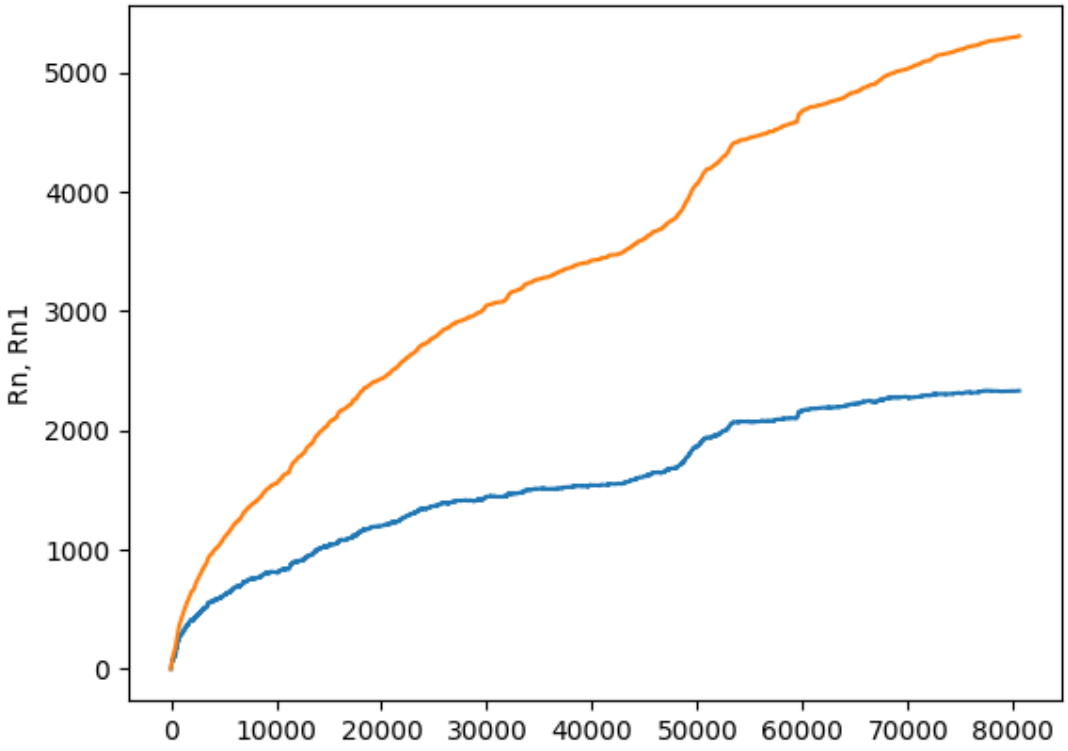

Figure 3: *The Picture of Dorian Gray* by Oscar Wilde

5 Conclusion

So, we study the number of words that occur exactly once since the beginning of the text as a stochastic process over the length of the text. The elementary probability model, going back to Bahadur and Karlin, states that the number of words that occur exactly once should grow according to a power law, like the number of different words. The final value of the number of words occurring exactly once is the number of hapaxes of this text. We construct two statistical tests to test Karlin's model under the assumption that the probabilities of words in this model satisfy the generalized Zipf's law. These statistical tests show that some texts fit the model well, but many texts deviate significantly from it. This deviation is that the number of hapaxes is too small relative to the number of different words. The longer the text, the more significant this deviation is. The elementary probability model predicts the number of hapaxes well for stories and novel chapters, but tends to be significantly wrong for long novels.

Acknowledgments

The research of Shahzod Fayzullaev is supported by the "El-yurt umidi" foundation under the Cabinet of Ministers of the Republic of Uzbekistan for training specialists abroad and communication with compatriots. The research of Artyom Kovalevskii is supported by the Fundamental scientific research of the SB RAS, project FWNF-2022-0010.

Authors are grateful to the anonymous referee for comments that allowed to significantly improve the presentation.

References

Abebe, B., Chebunin, M., Kovalevskii, A. (2023). Text segmentation via processes that count the number of different words forward and backward. *Journal of Quantitative Linguistics,* 31(1), pp. 1–18. https://doi.org/10.1080/09296174. 2023.2275342

Abebe, B., Chebunin, M., Kovalevskii, A., Zakrevskaya, N. (2022). Statistical tests for text homogeneity: using forward and backward processes of numbers of different words. *Glottometrics*, *53*, pp. 42–58. https://doi.org/10.53482/2022_53_401

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

Balasubrahmanyan, V., Naranan, S. (1996). Quantitative linguistics and complex system studies. *Journal of Quantitative Linguistics*, *3*(3), pp. 177–228. https://doi.org/10.1080/09296179608599629

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

Barbour, A., Gnedin, A. (2009). Small counts in the infinite occupancy scheme. *Electronic Journal of Probability*, *14*, Paper no. 13, pp. 365–384.

Ben-Hamou, A., Boucheron, S., Ohannessian, M. (2017). Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications. *Bernoulli*, *23*(1), pp. 249–287. https://doi.org/10.3150/15-BEJ743

Cancho, R., Solé, R. (2001). Two Regimes in the Frequency of Words and the Origins of Complex Lexicons: Zipf's Law Revisited. *Journal of Quantitative Linguistics*, *8*(3), pp. 165–173. https://doi.org/10.1076/jqul.8.3.165.4101

Chebunin, M., Kovalevskii, A. (2016). Functional central limit theorems for certain statistics in an infinite urn scheme. *Statistics and Probability Letters*, *119*, pp. 344–348. https://doi.org/10.1016/j.spl.2016.08.019

Chebunin, M., Kovalevskii, A. (2019). Asymptotically normal estimators for Zipf's law. *Sankhya A*, *81*, pp. 482–492. https://doi.org/10.1007/s13171-018-0135-9

Chebunin, M., Kovalevskii, A. (2022). Modifications of Karlin and Simon text models. *Siberian Electronic Mathematical Reports*, *19*, pp. 708–723. https://doi.org/10.33048/semi.2022.19.059

Decrouez, G., Grabchak, M., Paris, Q. (2018). Finite sample properties of the mean occupancy counts and probabilities. *Bernoulli*, *24*(3), pp. 1910–1941. https://doi.org/10.3150/16-BEJ915

Dutko, M. (1989). Central limit theorems for infinite urn models. *The Annals of Probability*, *17*, pp. 1255–1263. https://doi.org/ 10.1214/aop/1176991268

Gnedin, A., Hansen, B., Pitman, J. (2007). Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Probability Surveys*, *4*, pp. 146–171. https://doi.org/10.1214/07-PS092

Gnedin, A., Iksanov, A., Marynych, A. (2010). Limit theorems for the number of occupied boxes in the Bernoulli sieve. *Theory of Stochastic Processes*, *16(32)*, pp. 44–57.

Grabchak, M., Kelbert, M., Paris, Q. (2020). On the occupancy problem for a regime-switching model. *Journal of Applied Probability*, *57*(1), pp. 53–77. https://doi.org/10.1017/jpr.2020.33

Harrison, P. (1921). *The Problem of the Pastoral Epistles.* Oxford University Press.

Karlin, S. (1967). Central Limit Theorems for Certain Infinite Urn Schemes. *Jounal of Mathematics and Mechanics*, *17*(4), pp. 373–401.

Key, E. (1992). Rare Numbers. *Journal of Theoretical Probability*, *5*(2), pp. 375–389.

Manin, D. (2009). Mandelbrot's Model for Zipf's Law: Can Mandelbrot's Model Explain Zipf's Law for Language? *Journal of Quantitative Linguistics*, *16*(3), pp. 274–285. https://doi.org/10.1080/09296170902850358

Popescu, I.-I., Altmann, G. (2008). Hapax legomena and language typology. *Journal of Quantitative Linguistics*, *15*(4), pp. 370–378. https://doi.org/10.1080/09296170802326699

Tuzzi, A., Popescu, I.-I., Altmann, G. (2009). Zipf's Laws in Italian Texts. *Journal of Quantitative Linguistics*, *16*(4), pp. 354–367. https://doi.org/10.1080/09296170903211519

van Egmond, M., van Ewijk, L., Avrutin, S. (2015). Zipf's Law in Non-Fluent Aphasia. *Journal of Quantitative Linguistics*, *22*(3), pp. 233–249. https://doi.org/10.1080/09296174.2015.1037158

Zakrevskaya, N., Kovalevskii, A. (2001). One-parameter probabilistic models of text statistics. *Sibirskii Zhurnal Industrial'noi Matematiki*, *4*(2), pp. 142–153.

Appendix. Proofs

Proof of Theorem 1

We transform the difference of fractions:

$$\frac{R_{n,1}}{R_n} - \frac{R_{\lfloor n/2\rfloor,1}}{R_{\lfloor n/2\rfloor}} = \frac{\frac{R_{n,1}-\mathbf{E}R_{n,1}}{\mathbf{E}R_n} + \frac{\mathbf{E}R_{n,1}}{\mathbf{E}R_n}}{\frac{R_n-\mathbf{E}R_n}{\mathbf{E}R_n} + 1} - \frac{\frac{R_{\lfloor n/2\rfloor,1}-\mathbf{E}R_{\lfloor n/2\rfloor,1}}{\mathbf{E}R_n} + \frac{\mathbf{E}R_{\lfloor n/2\rfloor,1}}{\mathbf{E}R_n}}{\frac{R_{\lfloor n/2\rfloor}-\mathbf{E}R_{\lfloor n/2\rfloor}}{\mathbf{E}R_n} + \frac{\mathbf{E}R_{\lfloor n/2\rfloor}}{\mathbf{E}R_n}}$$

$$= \frac{\frac{Y_{n,1}(1)}{\sqrt{\mathbf{E}R_n}} + \frac{\mathbf{E}R_{n,1}}{\mathbf{E}R_n}}{\frac{Y_n(1)}{\sqrt{\mathbf{E}R_n}} + 1} - \frac{\frac{Y_{n,1}(1/2)}{\sqrt{\mathbf{E}R_n}} + \frac{\mathbf{E}R_{\lfloor n/2\rfloor,1}}{\mathbf{E}R_n}}{\frac{Y_n(1/2)}{\sqrt{\mathbf{E}R_n}} + \frac{\mathbf{E}R_{\lfloor n/2\rfloor}}{\mathbf{E}R_n}}.$$

By virtue of Lemmas 1 and 2 of Chebunin and Kovalevskii (2019), under the conditions of the theorem

$$\mathbf{E}R_n = \Gamma(1-\theta)c^\theta n^\theta + o(n^{\theta/2}),$$

$$\mathbf{E}R_{n,1} = \theta\Gamma(1-\theta)c^\theta n^\theta + o(n^{\theta/2}),$$

therefore

$$\mathbf{E}R_{n,1}/\mathbf{E}R_n = \theta + o(n^{-\theta/2}),$$

$$\mathbf{E}R_{[nt],1}/\mathbf{E}R_{n,1} = t^\theta + o(n^{-\theta/2}),$$

$$\mathbf{E}R_{[nt]}/\mathbf{E}R_n = t^\theta + o(n^{-\theta/2})$$

for $n \to \infty$, $t > 0$, and hence

$$\frac{R_{n,1}}{R_n} - \frac{R_{\lfloor n/2\rfloor,1}}{R_{\lfloor n/2\rfloor}} = \frac{\frac{Y_{n,1}(1)}{\sqrt{\mathbf{E}R_n}} + \theta}{\frac{Y_n(1)}{\sqrt{\mathbf{E}R_n}} + 1} - \frac{\frac{Y_{n,1}(1/2)}{\sqrt{\mathbf{E}R_n}} + \theta 2^{-\theta}}{\frac{Y_n(1/2)}{\sqrt{\mathbf{E}R_n}} + 2^{-\theta}} + o(n^{-\theta/2}).$$

If a sequence of random variables ξ_n converges in probability to zero, then

$$\frac{1}{1+\xi_n} = 1 - \xi_n + o_p\left(\xi_n\right).$$

Therefore

$$\begin{aligned}\frac{R_{n,1}}{R_n} - \frac{R_{\lfloor n/2\rfloor,1}}{R_{\lfloor n/2\rfloor}} &= \left(\frac{Y_{n,1}(1)}{\sqrt{\mathbf{E}R_n}} + \theta\right)\left(1 - \frac{Y_n(1)}{\sqrt{\mathbf{E}R_n}}\right) \\ &- \left(2^\theta\frac{Y_{n,1}(1/2)}{\sqrt{\mathbf{E}R_n}} + \theta\right)\left(1 - 2^\theta\frac{Y_n(1/2)}{\sqrt{\mathbf{E}R_n}}\right) + o_p(n^{-\theta/2}) \\ = \frac{1}{\sqrt{\mathbf{E}R_n}}&\left(Y_{n,1}(1) - \theta Y_n(1) - 2^\theta Y_{n,1}(1/2) + \theta 2^\theta Y_n(1/2)\right) + o_p(n^{-\theta/2}).\end{aligned}$$

Multiplying by $\sqrt{R_n}$, we obtain using the strong law of large numbers

$$\begin{aligned}&\sqrt{R_n}\left(\frac{R_{n,1}}{R_n} - \frac{R_{\lfloor n/2\rfloor,1}}{R_{\lfloor n/2\rfloor}}\right) \\ &= \frac{\sqrt{R_n}}{\sqrt{\mathbf{E}R_n}}\left(Y_{n,1}(1) - \theta Y_n(1) - 2^\theta Y_{n,1}(1/2) + \theta 2^\theta Y_n(1/2)\right) + o_p(1).\end{aligned}$$

By virtue of SLLN, $R_n/\mathbf{E}R_n$ converges to 1 a.s., and by virtue of the CLT $Y_{n,1}(1) - \theta Y_n(1) - 2^\theta Y_{n,1}(1/2) + \theta 2^\theta Y_n(1/2)$ converges weakly to a normal random variable $Y_1(1) - \theta Y(1) - 2^\theta Y_1(1/2) +$

$\theta 2^{\theta} Y(1/2)$ with zero expectation. Its variance is calculated based on the covariance matrix of the vector $(Y(1), Y(1/2), Y_1(1), Y_1(1/2))$, calculated in corollary 1.

The proof is complete.

Proof of Theorem 2

We transform the difference of fractions:

$$\frac{R_{\lfloor n/2\rfloor}}{R_n} - \frac{R_{\lfloor n/2\rfloor,1}}{R_{n,1}} = \frac{\frac{R_{\lfloor n/2\rfloor}-\mathbf{E}R_{\lfloor n/2\rfloor}}{\mathbf{E}R_n} + \frac{\mathbf{E}R_{\lfloor n/2\rfloor}}{\mathbf{E}R_n}}{\frac{R_n-\mathbf{E}R_n}{\mathbf{E}R_n} + 1} - \frac{\frac{R_{\lfloor n/2\rfloor,1}-\mathbf{E}R_{\lfloor n/2\rfloor,1}}{\mathbf{E}R_n} + \frac{\mathbf{E}R_{\lfloor n/2\rfloor,1}}{\mathbf{E}R_n}}{\frac{R_{n,1}-\mathbf{E}R_{n,1}}{\mathbf{E}R_n} + \frac{\mathbf{E}R_{n,1}}{\mathbf{E}R_n}}$$

$$= \frac{\frac{Y_n(1/2)}{\sqrt{\mathbf{E}R_n}} + \frac{\mathbf{E}R_{\lfloor n/2\rfloor}}{\mathbf{E}R_n}}{\frac{Y_n(1)}{\sqrt{\mathbf{E}R_n}} + 1} - \frac{\frac{Y_{n,1}(1/2)}{\sqrt{\mathbf{E}R_n}} + \frac{\mathbf{E}R_{\lfloor n/2\rfloor,1}}{\mathbf{E}R_n}}{\frac{Y_{n,1}(1)}{\sqrt{\mathbf{E}R_n}} + \frac{\mathbf{E}R_{n,1}}{\mathbf{E}R_n}}.$$

By virtue of Lemmas 1 and 2 of Chebunin and Kovalevskii (2019), under the conditions of the theorem

$$\mathbf{E}R_n = \Gamma(1-\theta)c^{\theta}n^{\theta} + o(n^{\theta/2}),$$

$$\mathbf{E}R_{n,1} = \theta\Gamma(1-\theta)c^{\theta}n^{\theta} + o(n^{\theta/2}),$$

therefore

$$\mathbf{E}R_{n,1}/\mathbf{E}R_n = \theta + o(n^{-\theta/2}),$$

$$\mathbf{E}R_{[nt],1}/\mathbf{E}R_{n,1} = t^{\theta} + o(n^{-\theta/2}),$$

$$\mathbf{E}R_{[nt]}/\mathbf{E}R_n = t^{\theta} + o(n^{-\theta/2})$$

for $n \to \infty$, $t > 0$, and hence

$$\frac{R_{\lfloor n/2\rfloor}}{R_n} - \frac{R_{\lfloor n/2\rfloor,1}}{R_{n,1}} = \frac{\frac{Y_n(1/2)}{\sqrt{\mathbf{E}R_n}} + 2^{-\theta}}{\frac{Y_n(1)}{\sqrt{\mathbf{E}R_n}} + 1} - \frac{\frac{Y_{n,1}(1/2)}{\sqrt{\mathbf{E}R_n}} + \theta 2^{-\theta}}{\frac{Y_{n,1}(1)}{\sqrt{\mathbf{E}R_n}} + \theta} + o(n^{-\theta/2}).$$

If a sequence of random variables ξ_n converges in probability to zero, then

$$\frac{1}{1+\xi_n} = 1 - \xi_n + o_p(\xi_n).$$

Therefore

$$\frac{R_{\lfloor n/2\rfloor}}{R_n} - \frac{R_{\lfloor n/2\rfloor,1}}{R_{n,1}} = \left(\frac{Y_n(1/2)}{\sqrt{\mathbf{E}R_n}} + 2^{-\theta}\right)\left(1 - \frac{Y_n(1)}{\sqrt{\mathbf{E}R_n}}\right)$$
$$- \left(\theta^{-1}\frac{Y_{n,1}(1/2)}{\sqrt{\mathbf{E}R_n}} + 2^{-\theta}\right)\left(1 - \frac{Y_{n,1}(1)}{\sqrt{\mathbf{E}R_n}}\right) + o_p(n^{-\theta/2})$$

$$= \frac{1}{\sqrt{\mathbf{E}R_n}} \left(Y_n(1/2) - 2^{-\theta} Y_n(1) - \theta^{-1} Y_{n,1}(1/2) + 2^{-\theta} Y_{n,1} \right) + o_p(n^{-\theta/2}).$$

Multiplying by $\sqrt{R_n}$, we obtain using the strong law of large numbers

$$\sqrt{R_n} \left(\frac{R_{\lfloor n/2 \rfloor}}{R_n} - \frac{R_{\lfloor n/2 \rfloor,1}}{R_{n,1}} \right)$$

$$= \frac{\sqrt{R_n}}{\sqrt{\mathbf{E}R_n}} \left(Y_n(1/2) - 2^{-\theta} Y_n(1) - \theta^{-1} Y_{n,1}(1/2) + 2^{-\theta} Y_{n,1} \right) + o_p(1).$$

By virtue of SLLN, $R_n/\mathbf{E}R_n$ converges to 1 a.s., and by virtue of the CLT $Y_{n,1}(1) - \theta Y_n(1) - 2^{\theta} Y_{n,1}(1/2) + \theta 2^{\theta} Y_n(1/2)$ converges weakly to a normal random variable $Y(1/2) - 2^{-\theta} Y(1) - \theta^{-1} Y_1(1/2) + 2^{-\theta} Y_1(1)$ with zero expectation. Its variance is calculated based on the covariance matrix of the vector $(Y(1), Y(1/2), Y_1(1), Y_1(1/2))$.

The proof is complete.