ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 24 МАТЕМАТИКИ Выпуск 3
2017

А. П. Ковалевский (Новосибирск, НГТУ, НГУ, НГУЭУ). **Оценивание параметра закона Ципфа–Мандельброта по последовательности количеств разных элементов выборки.**

Предполагается, что случайные величины $X_1, \dots, X_n$ образуют выборку со значениями на множестве натуральных чисел: $\mathbf{P}\{X_1 = i\} = p_i > 0,\ i \geqslant 1$.

Наблюдается число разных элементов выборки $R_j,\ 1 \leqslant j \leqslant n : R_j = \sum_{i=1}^{\infty} \mathbf{I}\{\exists k \leqslant j :\ X_k = i\}$. Это число непустых урн в бесконечной урновой схеме [1], [2].

Мандельброт [3] предложил следующую асимптотику вероятностей появления слов в художественном тексте:

$$p_i \sim Ci^{-1/\theta}, \qquad 0 < \theta < 1, \quad C > 0. \tag{1}$$

Это — модификация закона Ципфа [4]. Такой характер убывания вероятностей в предположениях бесконечной урновой схемы обеспечивает степенной рост числа разных элементов [1]: $\mathbf{E}\, R_n \sim C^{\theta}\, \Gamma(1-\theta)\, n^{\theta},\ n \to \infty$. Здесь $\Gamma(\cdot)$ — гамма-функция.

Карлин [2] доказал УЗБЧ и ЦПТ для R_n. Чебунин и Ковалевский [5] доказали функциональную ЦПТ для $Z_n(\cdot) = (R_{[n\cdot]} - \mathbf{E}\, R_{[n\cdot]})/\sqrt{\mathbf{E}\, R_n}$. Предельный процесс Z_θ — центрированный гауссовский с непрерывными п. н. траекториями и ковариационной функцией $K(s,t) = (s+t)^{\theta} - \max(s^{\theta}, t^{\theta})$. Отметим, что этот же предельный процесс возникает в [6] при изучении последовательности сумм индикаторов непустых урн со случайными знаками.

Закревская и Ковалевский [7] получили состоятельную оценку параметра θ в виде неявной функции от R_n при более жестком по сравнению с (1) условии $p_i = Ci^{-1/\theta}$, $i \geqslant 1$. В этой ситуации константа C является функцией от θ. Чебунин [8] доказал состоятельность оценки $\ln R_n / \ln n$ в ситуации более общей, чем (1). Недостаток этой оценки в том, что она не является асимптотически нормальной.

Для построения асимптотически нормальной оценки параметра θ предлагается суммировать логарифмы числа разных элементов с соответствующей весовой функцией. Пусть $\widetilde{\theta} = \int_0^1 a(t) \ln R_{[nt]}\, dt$, где функция $a(\cdot)$ кусочно непрерывна, $a(t) = 0$ для $t \in [0, \delta]$, где

$$\delta \in (0,1), \quad \int_0^1 a(t) \ln t\, dt = 1, \quad \int_0^1 a(t)\, dt = 0.$$

Теорема 1. *В условиях* (1) *оценка* $\widetilde{\theta}$ *сильно состоятельна.*

Теорема 2. *Если* $p_i = C\, i^{-1/\theta},\ i \geqslant i_0$, *где* $i_0 \geqslant 1$, *то* $\sqrt{\mathbf{E} R_n}\,(\widetilde{\theta} - \theta)$ *сходится слабо к случайной величине* $\int_0^1 a(t)\, t^{-\theta} Z_\theta(t)\, dt$, *имеющей нормальное распределение с нулевым математическим ожиданием.*

Автор благодарит М. Г. Чебунина за многочисленные полезные обсуждения. Исследование поддержано грантом РФФИ 17-01-00683.

СПИСОК ЛИТЕРАТУРЫ

1. *Bahadur R. R.* On the number of distinct values in a large sample from an infinite discrete distribution. — Proc. Nat. Inst. Sci. India, 1960, 26A, Supp. II, с. 67–75.
2. *Karlin S.* Central limit theorems for certain infinite urn schemes. — J. Math. Mech., 1967, v. 17, № 4, с. 373–401.
3. *Mandelbrot B.* Information theory and psycholinguistics: a theory of word frequencies. — In: Readings in Mathematical and Social Science. Cambridge: M.I.T. Press, 1968, с. 350–368.
4. *Zipf G. K.* Human behavior and the principle of least effort. Cambridge: Cambridge Univ. Press, 1949.
5. *Chebunin M.*, *Kovalevskii A.* Functional central limit theorems for certain statistics in an infinite urn scheme. — Statist. Probab. Lett., 2016, v. 119, с. 344–348.
6. *Durieu O.*, *Wang Y.* From infinite urn schemes to decompositions of self-similar Gaussian processes. — Electron. J. Probab., 2016, v. 21, paper No. 43, 23 с.
7. *Закревская Н. С.*, *Ковалевский А. П.* Однопараметрические вероятностные модели статистик текста. — Сиб. ж. индустр. матем., 2001, т. 4, в. 2, с. 142–153.
8. *Чебунин М. Г.* Оценивание параметров вероятностных моделей по числу различных элементов выборки. — Сиб. ж. индустр. матем., 2014, т. 17, в. 3, с. 135—147.