Секция «Теория вероятностей и математическая статистика»

# Предельный процесс для последовательных количеств слов, встретившихся в тексте ровно один раз.

**Научный руководитель – Ковалевский Артём Павлович**

***Файзуллаев Шахзод Шухрат угли***

*Аспирант*

Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

*E-mail: s.faizullaev@g.nsu.ru*

Гапаксы (hapax legomena) — это слова, которые встречаются в тексте (или корпусе текстов) только один раз. Харрисон [7] отметил, что количество hapax legomena в тексте характеризует авторский стиль. Однако была неясна правильная зависимость количества hapax legomena от количества всех слов в тексте.

Карлин [4] исследовал общую модель бесконечной урновой схемы: каждый шар независимо от других выбирает один из бесконечного множества ящиков в соответствии с некоторым распределением вероятностей (одинаковым для всех шаров).

Эту бесконечную урновую схему можно интерпретировать как элементарную вероятностную модель текста: ящики соответствуют словам бесконечного словаря, шарики соответствуют последовательным словам текста. Таким образом, слова в этой модели выбираются случайным образом и независимо друг от друга.

Карлин [4] изучил очень общий случай распределения вероятностей, обобщающий закон Ципфа.

Среди прочего он изучал статистику $R_n$ количества непустых урн после бросания $n$ шаров. Эта статистика соответствует количеству разных слов в элементарной вероятностной модели текста для текста из $n$ слов.

Мы изучаем количество слов, встречающихся ровно один раз с начала текста. Мы моделируем его как случайный процесс на протяжении всего текста. Элементарная вероятностная модель, восходящая к Бахадуру [6] и Карлину [4], утверждает, что количество слов, встречающихся ровно один раз, должно расти по степенному закону, как и количество различных слов. Итоговое значение количества слов, встречающихся ровно один раз, и есть количество гапаксов этого текста. Мы строим два статистических теста для проверки модели Карлина в предположении, что вероятности слов в этой модели удовлетворяют обобщенному закону Ципфа. Они основаны на следующих статистиках:

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

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

где $R_n$ - число разных слов во всем тексте, $R_{\lfloor n/2\rfloor}$- число разных слов в первой половине текста, $R_{n,1}$ - число слов, встретившихся ровно один раз во всем тексте (гапаксов этого текста), $R_{\lfloor n/2\rfloor,1}$ - число слов, встретившихся ровно один раз в первой половине текста.

Эти статистические тесты показывают, что некоторые тексты хорошо соответствуют модели, но многие тексты значительно от нее отклоняются. Это отклонение заключается в том, что количество гапаксов слишком мало по сравнению с количеством разных слов.

Карлин [4] доказал, в частности, усиленный закон больших чисел и центральную предельную теорему для $R_{n,1}$.

Кей [5] изучил асимптотику числа уникальных слов. В частности, показано, что если $\lim_{i\to\infty} p_{i+1}/p_i = 1$, то $R_{n,1} \to_p \infty$ при $n \to \infty$ . Если $\limsup_{i\to\infty} p_{i+1}/p_i lt; 1$, то $\mathbf{E}R_{n,1}$ равномерно ограничено.

Статистики $R_n$ и $R_{n,1}$ изучаются также в работах Бахадура [6], Чебунина и Ковалевского [2,3], Закревской и Ковалевского [1].

В данной работе доказана центральная предельная теорема для представленных выше статистик $V_n^{(1)}$ и $V_n^{(2)}$.

## Источники и литература


1) Н. С. Закревская, А. П. Ковалевский, 2001. Однопараметрические вероятностные модели статистик текста. Сиб. журн. индустр. матем., Т. 4, 2, с. 142–153.
2) M. Chebunin, A. Kovalevskii, *Functional central limit theorems for certain statistics in an infinite urn scheme*, Statistics and Probability Letters, **119** (2016), 344–348.
3) M. Chebunin, A. Kovalevskii, *Asymptotically normal estimators for Zipf's law*, Sankhya A (2019), **81**, 482–492.
4) S. Karlin, *Central Limit Theorems for Certain Infinite Urn Schemes*, Jounal of Mathematics and Mechanics, **17**, No. 4 (1967), 373–401. MR0216548
5) E. S. Key, 1992. Rare Numbers. Journal of Theoretical Probability, Vol. 5, No. 2, 375–389.
6) 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.
7) Harrison, P. (1921). The Problem of the Pastoral Epistles (O. U. Press, Ed.).