ОЦІНКА ДИСПЕРСІЇ ВИПАДКОВОЇ ВЕЛИЧИНИ ПО ХОДУ НАКОПИЧЕННЯ НЕЗАЛЕЖНИХ СПОСТЕРЕЖЕНЬ

Жуковська О.А.[1], Глушаускайте І.В.[1], Глушаускене Г.А [2]., Файнзільберг Л.С[2].

[1] Національний технічний університет України "Київський політехнічний інститут", 03056, м.Київ-56, пр. Перемоги, 37, тел. (044) 411-69-04, zhukovskaya@voliacable.com, tinweistime@gmail.com

[2] Міжнародний науково-навчальний центр інформаційних технологій та систем, 03680, Київ-680, пр. Академіка Глушкова, 40, тел. (044) 526-11-54

fainzilberg@voliacable.com, galinaglushauskene@rambler.ru

При вирішення прикладних задач часто необхідно оцінювати дисперсію випадкової величини за вібіркою незалежних спостережень. Відомо [1-4], що класичне співвідношення для обчислення незсуненої та спроможної оцінки дисперсії D випадкової величини X за скінченною вибіркою спостережень X, що становлять послідовність незалежних однаково розподілених величин $x_1,..., x_n$, має вигляд

$$\overline{D}_n = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{m})^2\,, \qquad \text{де } \overline{m} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

Проте обчислення дисперсії за такою формулою може привести до ряду проблем. Одна з істотних проблем при обробці реальних даних – неможливість приступити до обробки даних до здобуття всього масиву даних, що робить неможливим обчислення дисперсії в режимі он-лайн. Додаткові проблеми виникають через необхідність резервування комп'ютерної пам'яті для всього масиву даних, що при великій кількості спостережень обмежує практичне вживання цієї формули [3]. До того ж безпосереднє обчислення суми великої кількості значень може привести до переповнювання розрядної сітки [4].

Існують і рекурентні формули обчислення дисперсії [5], проте при великому об'ємі вибірки рекурентне обчислення приводить до відомої проблеми накопичення помилок округлення, що виникають на кожному кроці при виконанні математичних операцій.

Оскільки на практиці часто необхідне здобуття поточної оцінки дисперсії випадкової величини у міру накопичення даних, то замість (1) зручніше використовувати оцінку

$$\tilde{D}_n = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \tilde{m}_i)^2, \quad n = 2,3,\ldots, \tag{2}$$

де $\tilde{m}_i$ – поточна оцінка математичного сподівання M випадкової величини X, обчислена на i-му кроці за формулою

$$\tilde{m}_i = \frac{1}{i}\sum_{k=1}^{i} x_k \tag{3}$$

Але, як показано в роботі [6], оцінка (2), на відміну від традиційної оцінка (1) є зсуненою. Тому пропонується використовувати незсунену модифікацію (2):

$$\breve{D}_n = \frac{1}{n - \Gamma(n)}\sum_{i=1}^{n}(x_i - \tilde{m}_i)^2, \tag{4}$$

де $\Gamma(n)$ – часткова сума гармонійного ряду

$$\Gamma(n) = 1 + \frac{1}{2} + \frac{1}{3} + \ldots. + \frac{1}{n} = \ln n + \gamma_0 + o(1) \approx \ln n + \gamma_0, \tag{5}$$

а $\gamma_0 = 0{,}57721564\ 490\ldots$ – постійна Ейлера-Маскероні .

Для практичного застосування (4) важливо знати, чи є модифікована оцінка (4) спроможною, тобто чи збігається за ймовірністю послідовність оцінок $\breve{D}_n$ до величини D, що оцінюється. Традиційно перевірка спроможності оцінки проводиться з використанням закону великих чисел у будь-якій формі. Однак, очевидно, що величини під знаком суми виразу (4), тобто $(x_i - \tilde{m}_i)^2$ та $(x_{i+1} - \tilde{m}_{i+1})^2$, не є незалежними, а всі класичні теореми про закон великих чисел розглядають послідовність незалежних величин, що стоять під знаком суми.

Отже, ми не в змозі застосувати закон великих чисел до нашої послідовності оцінок, тому вимушені обрати один з обхідних шляхів для знаходження границі за ймовірністю нашої послідовності оцінок.

Оскільки оцінка $\overline{D}_n$ спроможна, то для доведення спроможності $\breve{D}_n$ достатньо показати, що

$$\lim_{n\to\infty} \breve{D}_n = \lim_{n\to\infty} \overline{D}_n \,. \tag{6}$$

Це твердження можна довести за допомогою теореми Теплиця [6].

Більш того, оскільки стандартна оцінка $\overline{D}_n$ є сильно спроможною, то модифікована оцінка (4) також є сильно спроможною.

Таким чином, запропонована модифікована оцінка (4), яка є більш зручною для обчислення поточного значення дисперсії в міру накопичення даних, як і традиційна, є незміщеною та сильно спроможною.

1. Reza A. Soltani, Moeanaddin R. On Dispersion of Stable Random Vectors and Its Application in the Prediction of Multivariate Stable Processes // Journal of Applied Probability.– 1994.– Vol. 31.–No. 3.– P. 691-699.
2. Brick, J.M., Morganstein, D. WesVarPC: Software for Computing Variance Estimates from Complex Designs // Proceedings of the 1996 Annual Research Conference.– 1996: Washington, U.S. Bureau of the Census. – P. 861-866
3. Kagan A., Shepp L.A. Why the variance? // Statistics and Probability Letters. – 1998. – 38, N4. – P.329-333.
4. Wolter K.M. Introduction to variance estimation.– 2007.– New York : Springer Verlag.– 449 P.
5. Knuth D.E. The art of Computer Programming. Vol.2.Seminumerical Algorithms, 3rd edn. – Boston: Addison-Wesley, 1998. – 565 p.
6. Жуковська О.А., Глушаускене Г.А., Файнзільберг Л.С. Дослідження властивостей модифікованої оцінки дисперсії випадкової величини за вибіркою незалежних спостережень //Наукові вісті НТТУ «КПІ» - 2008 - № 4.