УДК 519.241

Кутковецький В.Я., д-р техн. наук, професор кафедри комп’ютерних технологій МФ НаУКМА;
Турти М.В. . Кандидат технічних наук;

Кутковецький Валентин Якович, 1936 р.н. Доктор технічних наук, професор кафедри комп’ютерних технологій МФ НаУКМА. Коло наукових інтересів: моделювання за допомогою ЕОМ процесів в електрообладнанні; напівпровідникові перетворювачі електричної енергії, теоретичні основи електротехніки.

Турти Марина Валентинівна, 1961 р.н. Кандидат технічних наук, дисертація на тему: “Оптимізація електромагнітних елементів високочастотних перетворювачів енергії”. Коло наукових інтересів: моделювання перехідних процесів в електротехнічних пристроях, методи оптимізації, методи обробки даних експериментів. Має 29 наукових праць, 3 винаходи.

Про недоцільність використання медіани Ходжеса-Лемана в однофакторному аналізі

В статті показується можливість суттєвого спрощення статистичних розрахунків. Отримана у загальному вигляді формула дозволяє замінити медіану Ходжеса-Лемана при скороченні проміжних розрахунків у три та більше разів.

In the article possibility of essential simplifying of statistical computations is showed. Formula, which proofed in the general form possibility of replacing Hodges-Leman's median with the arithmetic mean of the sample has been received. Such change reduces the number of simple calculation in three and more times.

Метою однофакторного аналізу є виявлення впливу одного фактору на кінцевий результат, який називають відгуком. Однофакторний аналіз широко використовується в різних галузях людської діяльності (в економіці, екології, психології, медицині, техниці та ін.), і тому його вдосконалення має практичне значення.

В статті показується можливість суттєвого спрощення статистичних розрахунків при обробці спостережень, отриманих за методом однофакторного аналізу.

Для демонстрації властивостей медіани Ходжеса-Лемана використаємо приклад, наведений у (табл. 1) [1].

Для виявлення впливу грошового стимулювання на продуктивність праці шістьом групам людей, кожна з яких складається з 5-ти чоловік, було запропоновано розв’язати задачі однакової складності (кожній людині задачі ставились окремо). Групи різняться за величиною грошової винагороди: більшому номеру групи відповідає більша винагорода.

В даному випадку ми не проводитимо перевірку належності даних до одного і того ж розподілу

Таблиця 1. Кількість розв’язаних задач

Кількість розв’язаних задач	**Номер групи**					
	1	2	3	4	5	6
Кількість задач, розв’язаних 1-им членом групи	10	8	12	12	24	19
Кількість задач, розв’язаних 2-им членом групи	11	10	17	15	16	18
Кількість задач, розв’язаних 3-ім членом групи	9	16	14	16	22	27
Кількість задач, розв’язаних 4-им членом групи	13	13	9	16	18	25
Кількість задач, розв’язаних 5-им членом групи	7	12	16	19	20	24

Таблиця 2. Розрахунок за медіаною Ходжеса-Лемана зміщення між 1-ою та 4-ою вибірками

Номер члена групи, t	Кількість задач, розв’язаних t-им членом групи		Результати розрахунків x_{u1} - x_{t4}				
	x_{t1}	x_{t4}	x_{11}-x_{t4}	x_{21}-x_{t4}	x_{31}-x_{t4}	x_{41}-x_{t4}	x_{51}-x_{t4}
1	10	12	10-12 = -2	11-12 = -1	9-12 = -3	13-12= +1	7-12 = -5
2	11	15	10-15 = -5	11-15 = -4	9-15 = -6	13-15 = -2	7-15 = -8
3	9	16	10-16 = -6	11-16 = -5	9-16 = -7	13-16 = -3	7-16 = -9
4	13	16	10-16 = -6	11-16 = -5	9-16 = -7	13-16 = -3	7-16 = -9
5	7	19	10-19 = -9	11-19 = -8	9-19 = -10	13-19 = -6	7-19 = -12
$\sum u = \sum_{t=1}^{5} x_{u1} - x_{t4}$			$\sum_1$=-28	$\sum_2$=-23	$\sum_3$=-33	$\sum_4$=-13	$\sum_5$=-43

(нульову гіпотезу), бо в [1] за допомогою статистики Краскела-Уолеса доведено, що грошова винагорода статистично впливає на продуктивність праці.

Але далі в [1] для розрахунку зміщення однієї вибірки відносно іншої пропонується використовувати медіану Ходжеса-Лемана

Зупинимось на практичному використанні медіани Ходжеса-Лемана для даних табл. 1. Медіана

Ходжеса-Лемана дає алгоритм для розрахунку

$$z_{ij} = \mathrm{med}(x_{ui} - x_{vi}, \quad u = 1,\dots, A_i, \quad v = 1,\dots, B_j).$$

величини зміщення одної вибірки по відношенню до іншої в умовах, коли кількості вимірювань в різних вибірках в загальному випадку не є однаковими. Наприклад, розрахуємо зміщення між 1-ою та 4-ою вибірками (група 1 та група 4, відповідно i=1; j=4) за медіаною Ходжеса-Лемана (табл. 2).

Загальна сума різниць дорівнює

Тоді для отримання зміщення між 1-ою та 4-ою вибірками потрібно суму Σ=-140 поділити на

$$\sum = \sum_{u=1}^{5} u = -28 - 23 - 33 - 13 - 43 = -140. \quad (1)$$

В результаті отримуємо величину зміщення між вибірками

$$N = A_i B_j = 5 \cdot 5 = 25. \quad (2)$$

Тепер розглянемо алгоритм Ходжеса-Лемана в загальному вигляді. Із наведеного прикладу випливає, що медіана Ходжеса-Лемана розраховується за формулою

$$z_{14} = \frac{\sum}{N} = -\frac{140}{25} = -5{,}6.$$

де i = 1... A_i – порядковий номер спостереження і-ої вибірки; j = 1... B_j – порядковий номер спостереження j-ої вибірки; A_i, B_j – кількість спостережень вибірок, що розглядаються; a_i, b_j – спостереження вибірок, що розглядяються.

$$M = \frac{1}{A_i B_j}\sum_{i=1}^{A_i}\sum_{j=1}^{B_j}(a_i - b_j) =$$

$$= \frac{1}{A_i B_j}\sum_{i=1}^{A_i}\sum_{j=1}^{B_j} a_i - \frac{1}{A_i B_j}\sum_{i=1}^{A_i}\sum_{j=1}^{B_j} b_i =$$

$$= \frac{1}{A_i B_j}\sum_{i=1}^{A_i} B_j a_i - \frac{1}{A_i B_j}\sum_{j=1}^{B_j} A_i b_j =$$

$$= \frac{1}{A_i}\sum_{i=1}^{A_i} a_i - \frac{1}{B_j}\sum_{j=1}^{B_j} b_j,$$

З формули (3) випливає, що медіана Ходжеса-Лемана дає результат, тотожний з різницею середніх значень вибірок, що розглядаються. Тобто, для табл. 1 можна виконати розрахунок:

$$z_{14} = \mathbf{\Sigma}_1 - \mathbf{\Sigma}_4 =$$
$$= (10 + 11 + 9 + 13 + 7)/5 -$$
$$- (12 + 15 + 16 + 19)/5 = 10 + 15{,}6 = -5{,}6.$$

Якщо виконати розрахунки для різних значень A_i та B_j, то можна впевнитись, що результати розрахунків по двох алгоритмах співпадають.

Оцінимо складність алгоритмів за медіаною Ходжеса-Лемана та за середніми значеннями. Для першого випадку потрібно виконати такі елементарні розрахунки:

- знайти A_iB_j різниць;
- знайти $(A_i - 1)\, B_j$ сум;
- знайти загальну суму (B_j – 1 елементарних розрахунків);
- розрахувати за (1) $N=A_iB_j$ (1 елементарний розрахунок);
- за (2) розрахувати зміщення (1 елементарний розрахунок).

Таким чином, загальна кількість елементарних розрахунків для першого випадку становить

$$F_1 = A_iB_j + (A_i - 1)\, B_j + (B_j - 1) + 1 + 1 =$$
$$= 2A_iB_j + 1. \quad (4)$$

Якщо знаходити зміщення за середніми значеннями вибірок, то потрібно:

- знайти $(A_i - 1)$ та $(A_i - 1)\, B_j + (B_j - 1)$ сум;
- обчислити середні значення для вибірок (2-ий елементарний розрахунок):
- знайти різницю середніх значень (1-ий елементарний розрахунок).

Загальна кількість елементарних розрахунків для цього алгоритму становить

$$F_2 = (A_i - 1) + (B_j - 1) + 2 + 1 =$$
$$= A_i + B_i + 1. \quad (5)$$

З (4) та (5) видно, що із зростанням кількості елементів у вибірках складність розрахунків зміщення між вибірками з використанням медіани Ходжеса-Лемана зростає значно швидше, ніж в разі використання алгоритму за середніми значеннями вибірок. Якщо розглянути відношення $F = F_1/F_2$, то легко показати, шо F є зростаючою функцією від A_i та B_j і має гіпотетичне мінімальне значення $F_{\min} = F$ (A_i = 1 , B_j = 1) = 1. Якщо кількість елементів у вибірках перевищує 4, то кількість елементарних розрахунків за першим алгоритмом буде більше ніж втричі перевищувати кількість елементарних розрахунків за запропонованим алгоритмом. Таким чином, немає ніякої потреби використовувати більш складний алгоритм Ходжеса-Лемана.

Висновки. **1.** Медіану Ходжеса-Лемана недоцільно використовувати в однофакторному аналізі, бо вона при ускладнених розрахунках дає результат, тотожний з різницею середніх значень вибірок, які розглядаються. Цей висновок відноситься як до вибірок, що мають однакову кількість спостережень, так і до вибірок з різною кількістю спостережень. **2.** Запропонований алгоритм є придатним для використання в однофакторному аналізі і практично не потребує додаткових витрат часу на вивчення, оскільки базується на відомих поняттях. **3.** Запропонований алгоритм дозволяє значно скоротити кількість розрахунків (в розглянутому випадку кількість розрахунків скорочена приблизно в 3 рази). Тому його можна рекомендувати для використання в однофакторному аналізі, особливо у випадках наявності великої кількості спостережень та рівнів факторів.

Література

1. Тюрин Ю.Н., Макаров А.А. Статистический анализ данных на компьютере. – М.: Инфра-М, 1998. – 528 с.