УДК 519.234.3+519.234.7

Член-кореспондент НАН України **С. І. Ляшко, В. В. Алєксєєнко, Д. А. Клюшин**

Непараметричний критерій однорідності двох вибірок на основі статистики включення

Розглянуто нову непараметричну статистику, основану на порядкових статистиках, що відображає близькість двох вибірок. Наведено алгоритм обчислення статистики. Побудовано критерій однорідності двох вибірок із довільним визначеним рівнем значущості.

Нехай $x = (x_1, x_2, \ldots, x_n)$ і $y = (y_1, y_2, \ldots, y_m)$ — вибірки, отримані простим випадковим вибором із генеральних сукупностей G та H відповідно, що мають абсолютно неперервні функції розподілу $F(u)$ і $G(u)$. Одна з основних задач статистичного розпізнавання образів — перевірка гіпотези H_0 про те, що вибірки x і y є однорідними, тобто $F(u) = G(u)$.

У роботі [1] розглядається статистика, що є сумою абсолютних відхилень між наближеними функціями щільності еталонної вибірки та еталонної вибірки, доповненої елементами тестової. Статистика має досить складну будову і не може застосовуватись для класифікації тестової вибірки серед класів, представлених еталонними вибірками різного обсягу.

Інша статистика, наведена у [2], будується як квадратичне відхилення між частотою і ймовірністю потрапляння елементів тестової вибірки в інтервали, визначені еталонною вибіркою. Для цієї статистики наводиться функція розподілу, але не наведений алгоритм обчислення, що ускладнює використання результатів на практиці.

На базі ідей цих двох робіт будується нова непараметрична статистика, що дозволяє розв'язувати задачу перевірки гіпотези H_0.

Статистика включення. Нехай $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$ — варіаційний ряд вибірки $x_1, x_2, \ldots, x_n$ і $y_{(1)} < y_{(2)} < \cdots < y_{(m)}$ — варіаційний ряд вибірки $y_1, y_2, \ldots, y_m$.

Згідно з гіпотезою Хілла,

$$P(x_k \in (x_{(i)}, x_{(j)})) = \frac{j-i}{n+1}. \tag{1}$$

Введемо позначення:

$$I_0 = (-\infty; x_{(1)}), \qquad I_n = (x_{(n)}, \infty), \qquad I_j = (x_{(j)}, x_{(j+1)}), \qquad j = \overline{1, n-1}. \tag{2}$$

Тоді, за формулою (1), має місце $p = P(x_k \in I_i)) = 1/(n+1)$.

Позначимо l_i кількість y_j, таких, що $y_j \in I_j$, і $f_i = l_i/m$ — частоту потрапляння елементів вибірки y_j в інтервали I_i.

Введемо статистику включення, яка є абсолютним відхиленням між частотою і ймовірністю потрапляння елементів вибірки y_j в інтервали I_i:

$$\eta = \sum_{i=0}^{n} |p - f_i|. \tag{3}$$

Введемо такі позначення:

$$\overline{b} = \begin{cases} \left[\dfrac{m}{n+1}\right] + 1, & \dfrac{m}{n+1} > \left[\dfrac{m}{n+1}\right] \\ \dfrac{m}{n+1}, & \dfrac{m}{n+1} = \left[\dfrac{m}{n+1}\right] \end{cases};$$

$\overline{h}$ — кількість i таких, що $l_i \geqslant \overline{b}$, $\overline{w} = \sum\limits_{l_i \geqslant \overline{b}} l_i - \overline{b}$, $\delta_+ = \overline{b} - \dfrac{m}{n+1}$; $\underline{b} = \overline{b} - 1$;

$\underline{h}$ — кількість i таких, що $l_i \leqslant \underline{b}$, $\underline{w} = \sum\limits_{l_i \leqslant \underline{b}} \underline{b} - l_i$, $\delta_- = \dfrac{m}{n+1} - \underline{b}$.

Лема 1. *$\underline{h}$, $\underline{w}$, δ_- можна визначити через $\overline{h}$, $\overline{w}$, δ_+.*
Доведення. $\underline{h} = n - \overline{h}$, де n — довжина вектора $\overline{l}$,

$$\underline{w} = \sum_{l_i \leqslant \underline{b}} \underline{b} - l_i = \sum_{l_i \leqslant \underline{b}} \underline{b} - \sum_{l_i \leqslant \underline{b}} l_i = \underline{hb} - \left(m - \sum_{l_i \geqslant \overline{b}} l_i \right) =$$

$$= (n - \overline{h})(\overline{b} - 1) - \left(m + \sum_{l_i \geqslant \overline{b}} (\overline{b} - l_i) - \sum_{l_i \geqslant \overline{b}} \overline{b} \right) =$$

$$= n\overline{b} - n - \overline{hb} + \overline{h} - (m + \overline{w} - \overline{hb}) = n\overline{b} + \overline{h} - n - m - \overline{w}.$$

Отже,

$$\delta_- = \frac{m}{n+1} - \underline{b} = \frac{m}{n+1} - \overline{b} + 1 = 1 - \delta_+.$$

Лема 2. *Пара $(\overline{h}, \overline{w})$ однозначно визначає статистику η для визначених m і n.*
Доведення.

$$\eta = \sum_{i=0}^{n} |p - f_i| = \sum_{l_i \geqslant \overline{b}} \frac{\left| \frac{m}{n+1} - l_i \right|}{m} + \sum_{l_i \leqslant \underline{b}} \frac{\left| \frac{m}{n+1} - l_i \right|}{m} = \frac{\sum\limits_{l_i \geqslant \overline{b}} \left(l_i - \frac{m}{n+1} \right) + \sum\limits_{l_i \leqslant \underline{b}} \left(\frac{m}{n+1} - l_i \right)}{m} =$$

$$= \frac{\sum\limits_{l_i \geqslant \overline{b}} (l_i - \overline{b} + \delta_+) + \sum\limits_{l_i \leqslant \underline{b}} (\underline{b} - l_i + \delta_-)}{m} = \frac{\overline{w} + \overline{h}\delta_+ + \underline{w} + \underline{h}\delta_-}{m}.$$

За лемою 1 маємо

$$\eta = \frac{\overline{w} + \overline{h}\delta_+ + n\overline{b} + \overline{h} - n - m - \overline{w} + (n - \overline{h})\delta_-}{m} = \eta(\overline{h}, \overline{w}).$$

Твердження леми справедливе, оскільки $\overline{b}$, δ_+ і δ_- обчислюються через m і n.

Розподіл статистики включення. Припустимо, що справедлива гіпотеза H_0, тобто $F(u) = G(u)$. Введемо в розгляд об'єднану вибірку $z_1, z_2, \ldots, z_{m+n} = x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_m$. Нехай $z_{(1)}, z_{(2)}, \ldots, z_{(m+n)}$ — відповідний варіаційний ряд.

Оскільки елементи вибірки y_j належать генеральній сукупності G і відповідному варіаційному ряду, то ми можемо вважати, що всі можливі позиції, що займають елементи y_j у варіаційному ряді, є рівноімовірними. Тобто будь-які m елементів вибірки $z_1, z_2, \ldots, z_{m+n}$ можуть належати тестовій вибірці. Набір позицій елементів тестової вибірки у варіаційному ряді однозначно визначає вектор $\overline{l}$ частот потрапляння елементів вибірки y_j в інтервали I_i. Всього різних розподілів $\overline{l}$ існує C^m_{m+n} — кількість способів обрати у варіаційному ряді m позицій, що відповідають елементам вибірки y_j.

Позначимо як $S(n,k,l)$ кількість способів розмістити k кульок по n урнах так, щоб в кожній було не більше, ніж l кульок. Ця величина може бути обчислена за допомогою таких рекурентних формул:

$$S(0,k,l)=0, \qquad S(n,0,l)=1, \qquad S(n,1,l)=n, \qquad S(1,k,l)=1, \qquad k\leqslant l,$$
$$S(n,k,l)=S(n-1,k,l)+S(n-1,k-1,l)+\cdots+S(n-1,k-l,l).$$

Теорема 1. *Для визначених характеристик $(\overline{h},\overline{w})$ кількість різних векторів $\overline{l}$, для яких характеристики будуть набувати цих значень, можна обчислити за формулою:*

$$N(\overline{h},\overline{w})=C^{\overline{h}}_{n+1}C^{\overline{w}}_{\overline{h}+\overline{w}-1}S(n+1-\overline{h},m-\overline{bh}-\overline{w},\overline{b}-1). \tag{4}$$

Доведення. Кількість способів обрати $\overline{h}$ позицій серед $n+1$ дорівнює $C^{\overline{h}}_{n+1}$. Згідно з позначеннями, $\overline{h}$ — кількість позицій вектора $\overline{l}$, вибіркові значення в яких перевищують або дорівнюють $\overline{b}$, $\overline{w}$ — величина, на яку за цими позиціями сума елементів в інтервалі більша за $\overline{b}$. Кількість векторів, в яких на $\overline{h}$ позиціях знаходяться числа, що перевищують або дорівнюють $\overline{b}$ так, що $\sum\limits_i (l_i-\overline{b})=\overline{w}$ — це те саме, що кількість способів спочатку розташувати $\overline{b}$ кульок у $\overline{h}$ урнах, потім ще додати $\overline{w}$ кульок до цих урн. Кількість способів — $C^{\overline{w}}_{\overline{h}+\overline{w}-1}$. На інших $n+1-\overline{h}$ позиціях залишається всього $m-\overline{bh}-\overline{w}$ елементів, причому кількість елементів в кожному інтервалі не більше ніж $\overline{b}-1$. Тобто кількість варіантів дорівнює $S(n+1-\overline{h},m-\overline{bh}-\overline{w},\overline{b}-1)$. Отже, загальна кількість векторів дорівнює $C^{\overline{h}}_{n+1}C^{\overline{w}}_{\overline{h}+\overline{w}-1}S(n+1-\overline{h},m-\overline{bh}-\overline{w},\overline{b}-1)$.

Таким чином, показано, що для визначених m і n імовірність того, що розподілу інтервалів $\overline{l}$ буде відповідати пара $(\overline{h},\overline{w})$, дорівнює

$$p(\overline{h},\overline{w})=\frac{C^{\overline{h}}_{n+1}C^{\overline{w}}_{\overline{h}+\overline{w}-1}S(n+1-\overline{h},m-\overline{bh}-\overline{w},\overline{b}-1)}{C^m_{m+n}}. \tag{5}$$

Критерії однорідності. На основі викладеного можна побудувати такий критерій для перевірки гіпотези H_0:

1) для даних m та n (розмірностей тестової та еталонної вибірки) обчислити ймовірності для всіх можливих пар (h,w). Зазначмо, що ці ймовірності можна обчислити заздалегідь і занести в таблицю, адже вони залежать тільки від m та n;

2) визначити пару характеристик (h_0,w_0) для заданих тестової і еталонної вибірки і відповідну ймовірність $p(h_0,w_0)$;

3) обчислити α — суму ймовірностей менш імовірних пар характеристик, ніж (h_0,w_0). Фактично α можемо вважати досяжним рівнем значущості, тобто найменшим рівнем значущості, при якому гіпотеза H_0 відхиляється для таких значень характеристик;

Таблиця 1. Розподіли еталонних вибірок

n	μ	σ
100	0	1
150	0	2
200	1	1
250	−1	4

Таблиця 2. Результати експерименту

Класифікація	Експеримент 1				Експеримент 2				Експеримент 3			
Розподіл	1	2	3	4	1	2	3	4	1	2	3	4
$К_{успіх}$	98	84	69	84	86	93	94	92	98	69	98	94
$К_{помилка}$	0	1	1	0	0	0	0	1	0	4	0	0
$К_{невизн}$	2	15	30	16	14	7	6	7	2	27	2	6
$АК_{успіх}$	99	96	93	99	98	100	100	99	99	94	100	99
$АК_{помилка}$	1	4	7	1	2	0	0	1	1	6	0	1

4) нехай бажаний рівень значущості α_0. Тоді критерієм прийняття гіпотези H_0 буде $\alpha > \alpha_0$.

Розглянемо задачу класифікації. Нехай задано k класів, кожен з яких представлений відповідною еталонною вибіркою $x^{(i)}$, $i = 1, \ldots, k$, і тестова вибірка y. Треба визначити, якому класу належить тестова вибірка. Для кожної еталонної вибірки і тестової вибірки за критерієм і заданим рівнем значущості перевіримо справедливість гіпотези H_0. Якщо гіпотеза H_0 приймається згідно з критерієм рівно для однієї еталонної вибірки, то вважаємо, що тестову вибірку вдалося класифікувати і вона належить класу цієї еталонної вибірки, інакше класифікацію виконати не вдалося.

Побудована згідно з алгоритмом величина $1 - \alpha$ може бути використана як міра близькості двох вибірок.

Якщо будь-що необхідно обрати якийсь із класів, то можна побудувати альтернативний критерій класифікації — будемо вважати, що тестова вибірка належить тому класу, для еталонної вибірки якого величина $1 - \alpha$ є мінімальною. В такому випадку невизначеність не виникає.

Практична перевірка критеріїв однорідності. Методика перевірки критеріїв класифікації має такий вигляд. Генеруємо чотири еталонні вибірки, які будуть представляти класи класифікації (табл. 1).

Для кожного типу нормального розподілу генеруємо 100 тестових вибірок по 300 елементів в кожній. Кожну тестову вибірку класифікуємо за критерієм (К) і альтернативним критерієм (АК). Експеримент повторимо тричі. Результати наведені у табл. 2.

1. *Алексєєнко В. В.* Статистика включення // Журн. обчисл. і прикл. математики. – 2012. – № 1(107). – 105–111 с.
2. *Bairamov I. G., Ozkaya N.* On the nonparametric test for two sample problems based on spacing // J. of App. Stat. Science. – 2000. – **10**, No 1. – P. 57–68.
3. *Hill B.* Posteriori distribution of percentiles: Bayes' theorem for sampling from a population // J. Amer. Statist. Assoc. – 1968. – **63**, No 322. – P. 677–691.

Київський національний університет ім. Тараса Шевченка

Надійшло до редакції 29.11.2012

Член-корреспондент НАН Украины **С. И. Ляшко, В. В. Алексеенко, Д. А. Клюшин**

Непараметрический критерий однородности двух выборок на основе статистики включения

Рассмотрена новая непараметрическая статистика, основанная на порядковых статистиках и отражающая близость двух выборок. Приведен алгоритм вычисления статистики. Построен критерий однородности двух выборок с произвольно заданным уровнем значимости.

Corresponding Member of the NAS of Ukraine **S. I. Lyashko, V. V. Alexeyenko, D. A. Klyushin**

Nonparametrical test for homogeneity of two samples based on the inclusion statistics

A new nonparametrical statistics based on order statistics reflecting the proximity of two samples is investigated. An algorithm for calculation of this statistics is proposed. The test for homogeneity of two samples with an arbitrary significance level is developed.