УДК 519.21

Р. Є. Майборода (Київ. нац. ун-т ім. Т. Шевченка)

КРИТЕРІЙ ОДНОРІДНОСТІ ДЛЯ СУМІШЕЙ ЗІ ЗМІННИМИ КОНЦЕНТРАЦІЯМИ

We construct the generalized Kolmogorov–Smirnov test for verifying a hypothesis on the homogeneity of sample against the alternative sample from a mixture with varying concentration. We obtain asymptotic formulae and nonasymptotic upper bounds for probabilities of errors of first and and second type.

Побудовано узагальнений критерій Колмогорова–Смірнова для перевірки гіпотези про однорідність вибірки проти альтернативи — вибірки з суміші зі змінними концентраціями. Отримано асимптотичні формули і неасимптотичні оцінки зверху для ймовірностей помилок першого і другого роду.

1. Вступ. Задачі аналізу сумішей з концентраціями, що змінюються, часто виникають при статистичній обробці медико-біологічних та соціально-психологічних даних. При цьому вважається, що спостережувані об'єкти $O_1, \dots, O_N$ можуть належати до однієї з M популяцій (генеральних сукупностей). Номер популяції, до якої належить об'єкт O, позначимо $\operatorname{ind}(O)$. Для кожного об'єкта O_j спостерігається деяка його (випадкова) числова характеристика ξ_j. Розподіл ξ_j може залежати від $\operatorname{ind}(O_j)$:

$$\Pr\{\xi_j < x \mid \operatorname{ind}(O_j) = k\} = H_k(x).$$

Функції $H_k(x)$ невідомі. Значення $\operatorname{ind}(O_j)$ теж невідоме, але відома ймовірність того, що об'єкт належить до даної популяції $w_k(j) = \Pr\{\operatorname{ind}(O_j) = k\}$ (концентрація об'єктів k-ї популяції в суміші, з якої був вибраний j-й об'єкт). Отже, спостереження — це набір незалежних випадкових величин (в. в.) $\xi_1, \dots, \xi_N$ з розподілами

$$\Pr\{\xi_j < x\} = F_j(x) = \sum_{j=1}^{M} w_k(j) H_k(x). \tag{1}$$

Задачі оцінки H_k розглядались у [1], класифікація даних за популяціями — у [2]. Дана робота присвячена побудові критерію для перевірки гіпотези $\mathbf{H}_0: H_1 = H_2 = \dots = H_M$ проти альтернативи $\mathbf{H}_1$: існують k, m, x такі, що $H_k(x) \neq H_m(x)$. $\mathbf{H}_0$ фактично, є гіпотезою про однаковий розподіл спостережень $\xi_1, \dots, \xi_N$. У п. 2 розглянуто загальну схему критерію, у п. 3 — його асимптотичні властивості, у п. 4 — неасимптотичні оцінки ймовірності помилок першого та другого роду.

2. Схема критерію. Будемо розглядати $w_k = (w_k(1), \dots, w_k(N))$ як елементи векторного простору R^N із скалярним добутком

$$\langle a, b \rangle_N = \frac{1}{N} \sum_{j=1}^{N} a(j) b(j).$$

Лінійний простір, натягнутий на w_k, $k = 1, \dots, M$, позначимо через W. Надалі вважаємо, що виконується умова:

А) w_k, $k = 1, \dots, M$, є лінійно незалежними в R^N.

Нехай $u_0 = (1, 1, \dots, 1)$, $u_1, \dots, u_{M-1}$ — ортонормований базис простору W ($u_0 \in W$, оскільки, за означенням, $\sum_{k=1}^{M} w_k(j) = 1$ для всіх j).

© Р. Є. МАЙБОРОДА, 2000

Введемо у розгляд зважені емпіричні функції розподілу

$$\hat{F}_N(x, u) = \frac{1}{N}\sum_{j=1}^{N} u(j) I\{\xi_j < x\},$$

де $I\{A\}$ — індикатор події A. Як статистику критерію однорідності використаємо величину

$$D_N = \sqrt{N} \sup_{k=1,\dots,M-1} \sup_{x\in R} \left|\hat{F}_N(x, u_k)\right|. \tag{2}$$

(Підкреслимо, що супремум береться по $k \geq 1$, тобто звичайна емпірична функція розподілу $\hat{F}_N(x, u_0)$ не враховується.) Задамо деякий поріг критерію $c > 0$. Гіпотеза $\mathbf{H}_0$ приймається, якщо $D_N \leq c$, і відхиляється, якщо $D_N > c$.

Запропонований критерій є узагальненням критерію однорідності Колмогорова – Смірнова (КСК) на випадок суміші зі змінними концентраціями. Дійсно, якщо в (1) покласти $M = 2$,

$$w_1(j) = \begin{cases} 1, & \text{якщо } j < N_1; \\ 0, & \text{якщо } j > N_2, \end{cases} \qquad w_2(j) = 1 - w_1(j),$$

то отримаємо звичайну задачу перевірки однорідності двох вибірок обсягів N_1 та N_2, для якої статистика КСК має вигляд

$$\mathrm{KC}_N = \sqrt{\frac{N_1(N-N_1)}{N}} \sup_x \left| \frac{1}{N_1}\sum_{j=1}^{N_1} I\{\xi_j < x\} - \frac{1}{N-N_1}\sum_{j=N_1+1}^{N} I\{\xi_j < x\} \right|.$$

Якщо в (2) покласти $u_1(j) = \dfrac{N}{\sqrt{N_1(N-N_1)}}\left(w_1(j) - \dfrac{N_1}{N}\right)$, то $\mathrm{KC}_N = D_N$.

Описаний критерій назвемо узагальненим КСК (УКСК), а функції $u_1, \dots, u_{M-1}$ — базисними функціями критерію. Відмітимо, що при $M > 2$ базисні функції визначені неоднозначно.

3. Асимптотичні властивості критерію. Для того щоб дослідити поведінку критерію при зростанні об'єму вибірки, розглянемо дані $\xi_1, \dots, \xi_N$ як елемент схеми серій $\{\xi_j^N\}$, $j = 1, \dots, N$, $N = 1, 2, \dots$, де ξ_j^N незалежні при фіксованому N,

$$\Pr\{\xi_j^N < x\} = \sum_{j=1}^{M} w_k^N(j) H_k^N(x) \tag{3}$$

(тобто функції розподілу H_k та концентрації w_k в (1) будемо вважати залежними від об'єму вибірки N). Відповідно, базисні функції УКСК будемо позначати $u_1^N, \dots, u_{M-1}^N$. Нехай $\alpha_N(c) = \Pr\{D_N > c \mid \mathbf{H}_0\}$ — ймовірність помилки першого роду при використанні УКСК з порогом c.

Теорема 1. *Нехай* $H^N = H_1^N = \dots = H_M^N$ *— неперервна функція при всіх* N,

$$\sup_{k=1,\dots,M-1} \sup_N \sup_{1\leq j\leq N} \left|u_k^N(j)\right| < C < \infty. \tag{4}$$

Тоді $\alpha_N(c)$ *не залежить від* H^N *і*

$$\lim_{n\to\infty} \alpha_N(c) = 1 - (K(c))^{M-1}, \tag{5}$$

де $K(c)$ *— функція розподілу Колмогорова.*

***Зауваження* 1.** При $M > 2$ вектори u_m^N задаються неоднозначно. Їх можна отримати, наприклад, ортогоналізацією Грама – Шмідта системи векторів

$W' = (u_0,\ w_1^N, \dots, w_{M-1}^N)$. В такому випадку умова (4) виконується, якщо $\det \Gamma_N > c > 0$ для всіх N, де Γ_N — матриця Грама системи векторів W' у скалярному добутку $\langle \cdot, \cdot \rangle_N$.

Доведення теореми. Незалежність $\alpha_N(c)$ від H^N доводиться з використанням квантильного перетворення так само, як і для звичайного КСК. Рівність (5) досить довести для випадку, коли H^N — рівномірний розподіл на $[0, 1]$. Покажемо, що в цьому випадку векторнозначний випадковий процес

$$Y_N(x) = \left(Y_N^1(x), \dots, Y_N^{M-1}(x)\right) = \left(\sqrt{N}\,\hat{F}_N(x, u_1), \dots, \sqrt{N}\,\hat{F}_N(x, u_{M-1})\right)$$

при $N \to \infty$ слабко збігається в $D[0, 1]$ до процесу

$$B(x) = (B_1(x), \dots, B_{M-1}(x)), \tag{6}$$

де $B_i(x)$ — незалежні між собою версії стандартного броунівського моста. Для цього підрахуємо

$$\mathbf{E}\, Y_N^k(x) = \frac{1}{N}\sum_{j=1}^{N} u_k^N(j) H^N(x) = 0 \tag{7}$$

(внаслідок ортогональності u_k^N та u_0). При $0 \le x, y \le 1$

$$\operatorname{cov}_{km}(x, y) =$$

$$= \mathbf{E}\, Y_N^k(x) Y_N^m(y) = \frac{1}{N}\sum_{j=1}^{N} u_k^N(j) u_m^N(j) \left(\mathbf{E} I\left\{\xi_j^N < x\right\} I\left\{\xi_j^N < y\right\} - xy\right) =$$

$$= I\{k = m\}(\min(x, y) - xy) = \mathbf{E} B_k(x) B_m(y) \tag{8}$$

(тут врахована ортогональність базисних функцій). Y_N є нормованою сумою незалежних доданків, кожний з яких, згідно з (4), обмежений за модулем константою C. Тому, використовуючи (7), (8), за центральною граничною теоремою отримуємо збіжність скінченновимірних розподілів Y_N до B. Використовуючи стандартну техніку [3], отримуємо оцінку

$$\mathbf{E}\left(Y_N^i(x) - Y_N^i(y)\right)^2 \left(Y_N^i(y) - Y_N^i(z)\right)^2 \le C_1 (z - x)^2$$

для деякої константи $C_1 < \infty$ та будь-яких $0 \le x < y < z \le 1$, $i = 1, \dots, M-1$. Звідси випливає слабка збіжність в $D[0, 1]$. Враховуючи, що

$$\Pr\left\{\sup_{k=1,\dots,M} \sup_{x \in [0,1]} |B_k(x)| > c\right\} = 1 - (K(c))^{M-1},$$

отримуємо (5).

Розглянемо похибку другого роду для даного критерію у випадку альтернатив, що зближуються. Для цього покладемо

$$\tilde{H}_k^N(x) = H(x) + \frac{h_k(x)}{\sqrt{N}}, \tag{9}$$

де H — неперервна функція розподілу, $h_k(x)$ — неперервні функції, такі, що $\tilde{H}_k^N(\cdot)$ є функцією розподілу при достатньо великих N.

Зауваження 2. Такий вибір альтернатив є загальноприйнятим для задачі аналізу критеріїв при альтернативах, що зближуються. Якщо, наприклад, $H(x)$ має щільність f_H, то як h_k можна вибрати будь-які функції вигляду $h_k(x) = \int_{-\infty}^{x} f_k(y)dy$, де $f_k \in L_1(R)$ такі, що $\int_{-\infty}^{\infty} f_k(y)dy = 0$, $f_k(x)/\sqrt{N_0} > -h_k(x)$. Тоді $\tilde{H}_k^N$ будуть функціями розподілу при $N > N_0$.

Позначимо $h = (h_1(\cdot), \dots, h_M(\cdot))$,

$$\beta_N(c, h) = \Pr\left\{D_N < c \,\middle|\, \tilde{H}_k^N(\cdot) = \tilde{H}_k^N(\cdot)\right\}$$

— ймовірність помилки другого роду при альтернативі, що має вигляд (9) з фіксованим h.

Теорема 2. *Нехай*:

1) *функції* $\tilde{H}_k^N$ *задаються формулою* (9);

2) *виконується співвідношення* (4);

3) *існують і є скінченними границі* $U_{ki} = \lim_{N\to\infty} \left\langle w_k^N, u_i^N \right\rangle_N$.

Тоді

$$\lim_{N\to\infty} \beta_N(c, h) = \Pr\left\{ \sup_{x\in R} \sup_{k=1,\dots,M-1} \left| \sum_{k=1}^{M} U_{ki} h_k(x) - B_i(H(x)) \right| < c \right\},$$

де B_i *— незалежні між собою версії стандартного броунівського моста.*

***Зауваження* 3.** Умови 2 і 3 теореми виконуються у двох важливих випадках:

1. Якщо $w_k^N(j) = \tilde{w}_k(j/N)$, де $\tilde{w}_k$ — деякі інтегровні за Ріманом функції на [0, 1], і виконується умова

$$\det \Gamma > 0, \tag{10}$$

де Γ — матриця Грама системи функцій $u_0 = 1$, $\tilde{w}_k$, $k = 1, \dots, M-1$, з скалярним добутком $\langle a, b \rangle_\infty = \int_0^1 a(t)b(t)\,dt$. Дійсно, в цьому випадку $\left\langle w_k^N, w_l^N \right\rangle_N \to$ $\to \langle \tilde{w}_k, \tilde{w}_l \rangle_\infty$.

2. Якщо $(w_k(j))_{k=1}^M$ є випадковими векторами, незалежними і однаково розподіленими при різних значеннях j, причому умова (10) виконується для матриці Грама з скалярним добутком $\langle w_k, w_l \rangle = \mathbf{E} w_k(1) w_l(1)$. В цьому випадку умова 3 виконується майже напевне, а незалежність $\xi_1, \dots, \xi_j$ та рівність (1) слід розуміти як умовні при фіксованих $w_k(j)$.

Доведення теореми. Будемо вважати, що $H_k^N = \tilde{H}_k^N$. Тоді

$$\mathbf{E}\sqrt{N}\,\hat{F}^N(x, u_i) = \frac{1}{N} \sum_{j=1}^{N} u_k^N(j) \sum_{k=1}^{M} w_j(j) h_k(x) \to \sum_{k=1}^{M} U_{ki} h_k(x)$$

при $N \to \infty$, рівномірно по x (внаслідок обмеженості $h(x)$). Аналогічно доведенню теореми 1 можна показати, що розподіл вектор-функції

$$Y_N(x) = \left(\sqrt{N}\left(\mathbf{E}\hat{F}^N(x, u_i) - \hat{F}^N(x, u_i)\right)\right)_{i=1}^{M-1}$$

слабко збігається в $D(R)$ до $B(H(x))$, де B визначено за формулою (6). Звідси отримуємо твердження теореми.

4. Неасимптотичні властивості критерію. Побудуємо тепер оцінку зверху для α_N при фіксованому N, незалежну від функції розподілу спостережень H. Для цього позначимо

$$Z = \max_{k=1,\dots,M-1} \left(\sum_{j=1}^{N-1} |u_k(j) - u_k(j+1)| + |u_k(N)| \right).$$

Теорема 3. *Якщо виконана основна гіпотеза* $\mathbf{H_0}$, *то для будь-якого* c *такого, що* $cN > Z$,

$$\left|\sum_{j=1}^{N} u_k(j)E_j(x)\right| > \sqrt{N}\,c, \tag{14}$$

то для будь-яких k та x

$$\Pr\{D_N < c\} \le \Pr\left\{\left|\sum_{j=1}^{N} u_k(j)\big(I\{\xi_j<x\} - E_j(x)\big) + \sum_{j=1}^{N} u_k(j)E_j(x)\right| \le \sqrt{N}\,c\right\} \le$$

$$\le \Pr\left\{\pm\sum_{j=1}^{N} u_k(j)\big(I\{\xi_j<x\} - E_j(x)\big) > \left|\sum_{j=1}^{N} u_k(j)E_j(x)\right| - \sqrt{N}\,c\right\},$$

де знак $\pm$ збігається зі знаком величини $\sum_{j=1}^{N} u_k(j)E_j(x)$.

Зауважимо, що

$$\sum_{j=1}^{N} u_k(j)E_j(x) = N\sum_{k} \langle u_k, w_l\rangle_N H_l(x).$$

Виберемо k та x, при яких досягається максимум в означенні V. Тоді (14) виконується за умовою теореми, і , отже, $\Pr\{D_N < c\} \le \Pr\left\{\sum_{j=1}^{N}\eta_j > NV - \sqrt{N}\,c\right\}$, де $\eta_j = \pm\big(I\{\xi_j<x\} - E_j(x)\big)u_k(j)$. Зауважимо, що η_j — незалежні субгауссові випадкові величини з субгауссовими штандартами $\|\eta_j\|_{sub} = |u_k(j)|$ (див. [5, 6]). Отже,

$$\Pr\left\{\sum_{j=1}^{N}\eta_j > \lambda\right\} \le \exp\left(-\frac{\lambda^2}{2\sum_{j=1}^{N}\|\eta_j\|_{sub}^2}\right),$$

звідки маємо

$$\Pr\{D_N < C\} \le -\frac{(NV - \sqrt{N}\,C)^2}{2\sum_{j=1}^{N}|u_k(j)|^2}.$$

Враховуючи нормованість u_k, отримуємо твердження теореми.

***Зауваження* 4.** Як відомо [7], індекс Ходжеса–Лемана (ефективність критерію за Ходжесом–Леманом) визначається як

$$\mathrm{eff}_{HL} = \lim_{N\to\infty} \frac{-2\ln\beta_N(c)}{N}$$

(чим менше індекс, тим ефективніший критерій). З (13) випливає, що для УКСК $\mathrm{eff}_{HL} \le (\limsup_{N\to\infty} V_N)^2$.

1. *Майборода Р. Є.* Оцінка розподілів компонентів сумішей з концентраціями, що змінюються // Укр. мат. журн. – 1996. – **48**, № 8. – С. 562–566.
2. *Maiboroda R. E.* A classification problem for mixtures with time-dependent consentrations // Proc. II Ukrainian-Hungrian Conf. „New Trends in Probab. Theory add Math. Statist." – 1993. – P. 125–134.
3. *Биллингсли П.* Сходимость вероятностных мер. – М.: Наука, 1977. – 352 с.
4. *Вапник В. Н., Червоненкис А. Я.* Теория распознавания образов. – М.: Наука, 1974. – 416 с.
5. *Булдыгин В. В., Козаченко Ю. В.* О субгауссовских случайных величинах // Укр. мат. журн. – 1980. – **32**, № 6. – С. 723–730.
6. *Островский Е. И.* Экспоненциальные оценки распределений максимума негауссовского случайного поля // Теория вероятностей и ее применения. – 1990. – **35**, вып. 3. – С. 482–493.
7. *Боровков А. А., Могульский А. А.* Большие уклонения и проверка статистических гипотез. – Новосибирск: Наука, 1992. – 222 с.

Одержано 03.09.98