Теорія ймовірностей
та математична статистика
Вип. 2(101)/2019, с. 157–168

Teoriya Ĭmovirnosteĭ
ta Matematychna Statystyka
No. 2(101)/2019, pp. 157–168

УДК 519.21

ТЕСТИ ДЛЯ ГІПОТЕЗ ПРО КВАНТИЛІ РОЗПОДІЛІВ КОМПОНЕНТІВ СУМІШІ

Р. Є. МАЙБОРОДА, О. В. СУГАКОВА

Анотація. Розглядається задача перевірки статистичних гіпотез про квантилі різних розподілів компонентів суміші зі змінними концентраціями. Прикладами таких гіпотез є гіпотеза про однорідність медіан різних компонентів або гіпотеза про однорідність інтерквартильних розмахів. Описано оцінки для квантилів у моделі суміші зі змінними концентраціями, отримано умови їх асимптотичної нормальності. Побудовано асимптотичні довірчі еліпсоїди та тести для перевірки лінійних гіпотез на основі цих оцінок. Якість роботи запропонованих алгоритмів на вибірках скінченного обсягу перевірена за допомогою імітаційного моделювання.

Ключові слова і фрази. Модель суміші зі змінними концентраціями, квантилі, асимптотична нормальність, лінійна гіпотеза.

2010 *Mathematics Subject Classification.* Primary 62G08; Secondary 62G20.

1. Вступ

У прикладній статистиці квантилі часто використовують для опису розподілу даних. Наприклад, популярною характеристикою середнього положення у вибірці є вибіркова медіана, а характеристикою розкиду — інтерквартильний розмах, див. [13, с. 29], [12, с. 355]. На класичній діаграмі box-whishker plot («коробочка з вусами») прийнято відображати медіану і квартилі вибірки, [10, с. 30]. Якщо для опису неоднорідних даних застосовується кілька розподілів, природно виникає задача перевірки гіпотез про рівність їхніх квантилів. У випадку, коли спостереження безпомилково розбиваються на групи з однаковими розподілами всередині кожної групи, існує велика кількість параметричних та непараметричних багатовибіркових тестів для перевірки, наприклад, гіпотези про рівність медіан розподілів у всіх групах, таких як тест однорідності середніх в однофакторному дисперсійному аналізі, медіанний тест, тест Краскела–Веллса (Kruskal–Wallis test), див. [4, с. 260] та ін.

Дані медичних, генетичних, соціологічних досліджень часто являють собою суміш кількох компонентів із різними розподілами, які неможливо класифікувати безпомилково. Для аналізу таких даних застосовують моделі скінченних сумішей [8, 9]. Якщо у такій моделі вважати, що концентрації компонентів у суміші (імовірності змішування) є різними для різних спостережень, ці моделі стають моделями сумішей зі змінними концентраціями (СЗК).

У цій статті розглядається техніка побудови статистичних тестів для перевірки гіпотез про квантилі розподілів у випадку, коли дані описуються непараметричною моделлю СЗК. Тести для перевірки однорідності розподілів різних компонентів у моделі СЗК раніше розглядались у роботах [1, 7]. Для цензурованих даних тести однорідності розподілів розглядались у [11]. Тести для моментів розподілів компонентів побудовані у роботі [3]. Тести для квантилів для моделей СЗК, наскільки нам відомо, досі не розглядались.

Модель СЗК і оцінки для квантилів розподілів компонентів на основі цієї моделі описані у [6]. Там само доведена асимптотична нормальність таких оцінок окремого компонента суміші. Для побудови тестів, які порівнюють квантилі різних розподілів,

бажано мати твердження про асимптотичну нормальність вектора, складеного з оцінок для різних компонентів. У цій статті доведено відповідну теорему і, на її основі, будуються довірчі еліпсоїди для векторів, складених з квантилів різних компонентів суміші, і будуються тести для перевірки гіпотез про них. Опис моделі СЗК та оцінок для квантилів міститься у розділі 2. Теоремі про асимптотичну нормальність векторів таких оцінок присвячено розділ 3. Оцінки асимптотичної коваріаційної матриці для цих оцінок побудовані у п. 4. Довірчі еліпсоїди для векторів квантилів побудовані у розділі 5. Тести для перевірки гіпотез про квантилі розглянуті у розділі 6. Результати імітаційного моделювання представлені у розділі 7. Висновки містяться у розділі 8.

2. Модель суміші зі змінними концентраціями й оцінки квантилів

Нехай кожен спостережуваний об'єкт O належить одній з M популяцій (компонентів суміші). Позначимо $\kappa(O)$ номер компонента, якому належить O. Справжнє значення $\kappa(O)$ невідоме, але відомі ймовірності $p^{(k)} = \mathsf{P}\{\kappa(O) = k\}$. Спостерігається випадкова характеристика (змінна) об'єкта O, $\xi = \xi(O)$. Розподіл $\xi(O)$ залежить від того, якому компоненту належить O:

$$F^{(m)}(x) = \mathsf{P}\{\xi(O) < x \mid \kappa(O) = m\}.$$

Таким чином, розподіл спостережуваної характеристики ξ є сумішшю розподілів компонентів з імовірностями змішування (концентраціями компонентів у суміші) $p^{(m)}$:

$$\mathsf{P}\{\xi(O) < x\} = \sum_{m=1}^{M} p^{(m)} F^{(m)}(x). \tag{1}$$

Таку модель розподілу називають моделлю скінченної суміші. Ми розглянемо випадок, коли для різних спостережень концентрації компонентів можуть бути різними. Нехай спостерігаються об'єкти $O_1, \ldots, O_n$, зі спостережуваними характеристиками $\xi_j = \xi_{j;n} = \xi(O_j)$. Позначимо

$$p_j^{(m)} = p_{j;n}^{(m)} = \mathsf{P}\{\kappa(O_j) = m\}$$

— концентрація m-го компонента у суміші під час j-го спостереження (нижній індекс $;n$ вказує, що відповідна величина визначена для набору даних обсягу n, ми будемо вказувати цей індекс лише при вивченні випадку $n \to \infty$). Таким чином,

$$\mathsf{P}\{\xi_j < x\} = \sum_{m=1}^{M} p_j^{(m)} F^{(m)}(x). \tag{2}$$

Спостережувані величини $\xi_{j;n}$ вважаються незалежними при фіксованому n. Рівність (2) задає модель суміші зі змінними концентраціями. Ми розглядаємо непараметричний варіант цієї моделі, тобто вважаємо розподіли компонентів $F^{(m)}$ повністю невідомими.

Для оцінювання $F^{(m)}(x)$ можна використати навантажену емпіричну функцію розподілу (п. 2.1. у [6]):

$$\hat{F}^{(m)}(x) = \hat{F}_{;n}^{(m)}(x) = \frac{1}{n} \sum_{j=1}^{n} a_j^{(m)} \mathbf{1}\{\xi_j < x\}, \tag{3}$$

де $a_j^{(m)} = a_{j;n}^{(m)}$ — мінімаксні вагові коефіцієнти, що визначаються таким чином. Позначимо

$$\gamma_{ik} = \frac{1}{n} \sum_{j=1}^{n} p_j^{(i)} p_j^{(k)}, \mathbf{\Gamma}_{;n} = (\gamma_{ik})_{i,k=1}^{M}. \tag{4}$$

Припустимо, що $\det \mathbf{\Gamma}_{;n} \neq 0$. Позначимо $(\bar{\gamma}_{ik})_{i,k=1}^{M} = \mathbf{\Gamma}_{;n}^{-1}$. Тоді

$$a_j^{(m)} = \sum_{k=1}^{M} \bar{\gamma}_{km} p_j^{(k)}. \tag{5}$$

За досить широких умов $\hat{F}^{(m)}(x)$ є незміщеною, рівномірно (по $x \in \mathbb{R}$) консистентною оцінкою $F^{(m)}(x)$. Однак, оскільки деякі з вагових коефіцієнтів $a_j^{(m)}$, визначених (5), обов'язково повинні бути від'ємними, $\hat{F}^{(m)}(x)$ не є функціями розподілу для ймовірнісного розподілу.

Перейдемо тепер до означення квантилів та їх оцінок за допомогою навантажених емпіричних функцій. Грубо кажучи, квантиль $Q^F(\alpha)$ рівня α, $0 < \alpha < 1$, для функції розподілу F — це таке число q, що

$$F(q) = \alpha. \tag{6}$$

Якщо функція F є неперервною і строго зростаючою, то це рівняння має єдиний розв'язок відносно q і таке означення є строгим. У загальному випадку (6) може не мати розв'язків, або мати їх нескінченно багато. Тому використовують певні узагальнення цього означення, наприклад:

$$Q_-^F(\alpha) = \inf\{x \in \mathbb{R} \ : \ F(x) \geq \alpha\}. \tag{7}$$

Якщо рівняння (6) має єдиний розв'язок, це означення збігається з попереднім. Функцію $Q_-^F(\alpha)$ називають квантильною функцією. Якщо F — функція розподілу і розв'язків рівняння (6) нескінченно багато, $Q_-^F(\alpha)$ є їх точною нижньою границею. Тому цю величину природно назвати нижнім квантилем рівня α для F. Можна визначити також верхній квантиль

$$Q_+^F(\alpha) = \sup\{x \in \mathbb{R} \ : \ F(x) \leq \alpha\}, \tag{8}$$

який дорівнює верхній межі для розв'язків (6).

Далі ми завжди будемо вважати, що справжні функції розподілу компонентів $F^{(m)}(x)$ є неперервними і строго зростаючими в околі оцінюваних квантилів. Але для їх оцінок $\hat{F}^{(m)}(x)$ це не так. Тому

$$Q^{F^{(m)}}(\alpha) = Q_-^{F^{(m)}}(\alpha) = Q_+^{F^{(m)}}(\alpha), \text{ але } Q_-^{\hat{F}^{(m)}}(\alpha) \leq Q_+^{\hat{F}^{(m)}}(\alpha).$$

Як оцінку $Q^{F^{(m)}}(\alpha)$ можна взяти будь-яку статистику, що потрапляє в інтервал $[Q_-^{\hat{F}^{(m)}}(\alpha), Q_+^{\hat{F}^{(m)}}(\alpha)]$. Наприклад, це може бути середина цього інтервалу:

$$\hat{Q}^{(m)}(\alpha) = \frac{1}{2}(Q_-^{\hat{F}^{(m)}}(\alpha) + Q_+^{\hat{F}^{(m)}}(\alpha)).$$

У п. 3.4 [6] розглядаються більш акуратні оцінки, що використовують згладжування навантажених емпірчних функцій ламаними.

3. Асимптотична нормальність оцінок квантилів

Нехай d — деяке натуральне число, $\mathbf{k} = (k_1, k_2, \ldots, k_d)$, $k_i \in \{1, \ldots, M\}$ — набір можливих значень номерів компонентів суміші, $\beta = (\beta_1, \ldots, \beta_d)$, $\beta_i \in (0,1)$ — набір значень рівнів квантилів. Для $i = 1, \ldots, d$ позначимо $q_i = Q_-^{F^{(k_i)}}(\beta_i)$ — квантиль, що оцінюється, $\hat{q}_i = \hat{q}_{i;n} = Q_-^{\hat{F}_{;n}^{(k_i)}}(\beta_i)$ — оцінка квантиля,

$$\mathbf{q} = (q_1, \ldots, q_d)^T, \ \hat{\mathbf{q}}_{;n} = (\hat{q}_{1;n}, \ldots, \hat{q}_{d;n})^T. \tag{9}$$

Сформулюємо теорему про асимптотичну нормальність нормованої різниці

$$\zeta_{;n} = (\zeta_{;n}^1, \ldots, \zeta_{;n}^d)^T = \sqrt{n}(\hat{\mathbf{q}}_{;n} - \mathbf{q}).$$

Для цього нам знадобляться такі позначення та умови.

Позначимо

$$w_{;n}(i_1,i_2,m_1,m_2)=\frac{1}{n}\sum_{j=1}^{n}a_{j;n}^{(i_1)}a_{j;n}^{(i_2)}p_{j;n}^{(m_1)}p_{j;n}^{(m_2)},\ w_{;n}(i_1,i_2,m)=\frac{1}{n}\sum_{j=1}^{n}a_{j;n}^{(i_1)}a_{j;n}^{(i_2)}p_{j;n}^{(m)}.$$

Умова A. Для деяких $c_1>0$ і $n_0<\infty$, $\det\mathbf{\Gamma}_{;n}>c_1$ при всіх $n\ge n_0$.

Умова B. Існують границі

$$w(i_1,i_2,m_1,m_2)=\lim_{n\to\infty}w_{;n}(i_1,i_2,m_1,m_2),\ w(i_1,i_2,m)=\lim_{n\to\infty}w_{;n}(i_1,i_2,m).$$

Умова C. Для кожного $i=1,\dots,d$ існує такий відрізок $D_i=[q_i^-,q_i^+]$, що $q_i\in D_i$, $F^{(m)}$, $m=1,\dots,M$, є абсолютно неперервними на D_i з неперервними щільностями $f^{(m)}$ і для деякого $c_2>0$, $f^{(m)}(x)>c_2$ для всіх $x\in D_i$.

Для $i,j=1,\dots,d$, позначимо

$$v_{ij}=\sum_{m=1}^{M}w(k_i,k_j,m)F^{(m)}(\min(q_i,q_j))-\sum_{m,l=1}^{M}w(k_i,k_j,m,l)F^{(m)}(q_i)F^{(l)}(q_j),\qquad(10)$$

$$s_{ij}=\frac{v_{ij}}{f^{(k_i)}(q_i)f^{(k_j)}(q_j)},\ \mathbf{S}=(s_{ij})_{i,j=1}^{d}.\qquad(11)$$

Теорема 3.1. *Нехай виконані умови* ***A,B,C****. Тоді* $\zeta_{;n}\xrightarrow{\mathrm{W}}N(0,\mathbf{S})$ *при* $n\to\infty$.

У випадку $d=1$ це випливає з теореми 3.4.3 із [6]. Далі ми наведемо схему її доведення для випадку $d=2$ (випадок $d>2$ принципово не відрізняється від $d=2$, але вимагає введення великої кількості багатовимірних індексів).

Схема доведення теореми 3.1. Позначимо

$$\hat F_{+;n}^{(m)}(x)=\sup_{t\le x}\hat F_{;n}^{(m)}(t),$$

$$B_{;n}^{(m)}(x)=\sqrt{n}(\hat F_{;n}^{(m)}(x)-F^{(m)}(x)),\ B_{+;n}^{(m)}(x)=\sqrt{n}(\hat F_{+;n}^{(m)}(x)-F^{(m)}(x)),$$

$$B_{;n}(x_1,x_2)=(B_{;n}^{(k_1)}(x_1),B_{;n}^{(k_2)}(x_2)),\ B_{+;n}(x_1,x_2)=(B_{+;n}^{(k_1)}(x_1),B_{+;n}^{(k_2)}(x_2)).$$

Лема 3.1. *В умовах теореми 3.1 для всіх* $m=1,\dots,M$,

$$\sup_{x_1\in D_{k_1},x_2\in D_{k_2}}|B_{+;n}(x_1,x_2)-B_{;n}(x_1,x_2)|\to 0\ \textit{за ймовірністю при}\ n\to\infty.$$

Твердження леми безпосередньо випливає з теореми 2.3.1 [6].

Розглянемо гауссову випадкову функцію (процес) $B(x_1,x_2)=(B^1(x_1),B^2(x_2))$, $x_1\in D_{k_1}$, $x_2\in D_{k_2}$ таку, що для $i,j=1,2$, $x_1\in D_{k_i}$, $x_2\in D_{k_j}$,

$$\mathsf{E}\,B(x_1,x_2)=0,$$

$$\mathsf{E}\,B^i(x_1)B^j(x_2)=v(i,j,x_1,x_2)=\sum_{m=1}^{M}w(k_i,k_j,m)F^{(m)}(\min(x_1,x_2))-$$
$$-\sum_{m,l=1}^{M}w(k_i,k_j,m,l)F^{(m)}(x_1)F^{(l)}(x_2).$$

Позначимо $D=D_{k_1}\times D_{k_2}$.

Лема 3.2. *В умовах теореми 3.1 існує гауссова функція з неперервними траєкторіями на* D *з нульовим математичним сподіванням і коваріаційною функцією* $v(i,j,x_1,x_2)$.

Те, що $v(i,j,x_1,x_2)$ є коваріаційною функцією, випливає з доведення леми 3.3. Потраєкторна неперервність відповідного гауссового процесу доводиться з використанням критерія неперервності Дадлі так само, як у доведенні леми 2.5.1 [5].

Розглянемо простір $\mathbb{D}$, що складається з функцій неперервних зліва (у розумінні [2]), які відображають D у $\mathbb{R}^2$. В умовах теореми $B_{;n}$ і $B_{+,n}$ належать цьому простору при всіх n.

Лема 3.3. *В умовах теореми 3.1 послідовність $B_{+;n}$ при $n\to\infty$ слабко збігається до B у просторі $\mathbb{D}$ з рівномірною нормою.*

Доведення. Спочатку доведемо, що $B_{;n}$ прямує слабко до B. Помітимо, що

$$B_{;n}^{(k_i)}(x)=\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\eta_j(x),$$

де

$$\eta_{j;n}(x)=a_{j;n}^{(k_i)}(\mathbf{1}\{\xi_{j;n}<x\}-\sum_{m=1}^{M}p_{j;n}^{(m)}F^{(m)}(x))$$

(тут використана рівність $\sum_{j=1}^{n}a_{j;n}^{(k)}p_{j;n}^{(m)}=\mathbf{1}\{k=m\}$). Безпосереднім обчисленням отримуємо, що

$$\mathsf{E}\,\eta_{j;n}(x)=0,$$

$$\mathsf{E}\,B_{;n}^{i}(x_1)B_{;n}^{j}(x_2)=v_{;n}(i,j,x_1,x_2)=\sum_{m=1}^{M}w_{;n}(k_i,k_j,m)F^{(m)}(\min(x_1,x_2))-$$
$$-\sum_{m,l=1}^{M}w_{;n}(k_i,k_j,m,l)F^{(m)}(x_1)F^{(l)}(x_2).$$

Враховуючи умову (**B**), отримуємо, що коваріаційні функції процесу $B_{;n}$ прямують до коваріаційної функції B при $n\to\infty$. За умовою (**A**) усі $\eta_{j;n}$ рівномірно обмежені. Тому за центральною граничною теоремою, скінченновимірні розподіли $B_{;n}$ прямують до відповідних розподілів B. Компактність сім'ї мір, породжених процесами $B_{;n}$ у просторі $\mathbb{D}$ з метрикою Скорохода доводиться з використанням критерію Бікела — Вічури так само, як у лемі 2.5.1 [5]. Зі збіжності скінченновимірних розподілів та компактності випливає слабка збіжність $B_{;n}$ до B у метриці Скорохода. Оскільки, за лемою 3.2, граничний процес B має неперервні траєкторії, то збіжність у метриці Скорохода еквівалентна збіжності $B_{;n}$ до B у рівномірній метриці. Звідси, враховуючи лему 3.2, отримуємо твердження леми 3.3. □

Доведення теореми 3.1. Ми покажемо, що, для будь-яких $x_1,x_2\in\mathbb{R}$, при $n\to\infty$,

$$P_{;n}=\mathsf{P}\{\zeta_{;n}^1<x_1,\zeta_{;n}^2<x_2\}\to\mathsf{P}\{B^1(q_1)/f^{(k_1)}(q_1)<x_1,B^2(q_2)/f^{(k_2)}(q_2)<x_2\}.\quad(12)$$

Звідси випливає твердження теореми у випадку $d=2$.

Помітимо, що нерівність $Q^{\hat F_{;n}^{(m)}}(\beta)<x$ еквівалентна $\hat F_{+;n}^{(m)}(x)<\beta$.

Тому

$$P_{;n}=\mathsf{P}\left\{\hat F_{+;n}^{(k_1)}\left(q_1+\frac{x_1}{\sqrt{n}}\right)<\beta_1,\ \ \hat F_{+;n}^{(k_2)}\left(q_2+\frac{x_2}{\sqrt{n}}\right)<\beta_2\right\}=$$
$$=\mathsf{P}\left\{B_{+;n}^{(k_1)}\left(q_1+\frac{x_1}{\sqrt{n}}\right)<\sqrt{n}\left(\beta_1-F^{(k_1)}\left(q_1+\frac{x_1}{\sqrt{n}}\right)\right),\right.$$
$$\left.B_{+;n}^{(k_2)}\left(q_2+\frac{x_2}{\sqrt{n}}\right)<\sqrt{n}\left(\beta_2-F^{(k_2)}\left(q_2+\frac{x_2}{\sqrt{n}}\right)\right)\right\}.$$

Зауважимо, що при $n \to \infty$,

$$\beta_2 - F^{(k_i)}\left(q_i + \frac{x_i}{\sqrt{n}}\right) \sim \frac{f^{(k_i)}(q_i)x_i}{\sqrt{n}}.$$

Тому, враховуючи слабку збіжність $B_{;n}$ до B у рівномірній метриці і неперервність траєкторій B, отримуємо

$$P_{;n} \to \mathsf{P}\{B^1(q_1) < f^{(k_1)}(q_1)x_1,\ B^2(q_2) < f^{(k_2)}(q_2)x_2\}.$$

Це дає (12).

Оскільки $(B^1(q_1)/f^{(k_1)}(q_1), B^2(q_2)/f^{(k_2)}(q_2))$ є гауссовим випадковим вектором із нульовим математичним сподіванням і коваріаційною матрицею $\mathbf{S}$, з (12) випливає твердження теореми. □

Зауваження 3.1. Використовуючи ту ж техніку, можна довести, що теорема 3.1 буде виконуватись і при використанні оцінок $\hat{q}_{i;n} = Q_+^{\hat{F}_{;n}^{(k_i)}}(\beta_i)$, або $\hat{q}_{i;n} = Q^{\hat{F}_{;n}^{(k_i)}}(\beta_i)$.

4. Оцінювання асимптотичної коваріаційної матриці $\mathbf{S}$

Щоб використовувати теорему 3.1 для перевірки гіпотез про квантилі, нам буде потрібна оцінка для коваріаційної матриці граничного нормального розподілу $\mathbf{S}$ (матриця розсіювання оцінок $\hat{\mathbf{q}}_{;n}$). Для побудови оцінки замінимо в (11) теоретичні характеристики моделі їх емпіричними аналогами.

Величини v_{ij}, задані (10), можна оцінити за допомогою

$$\begin{aligned}\hat{v}_{ij;n} = &\sum_{m=1}^{M} w_{j;n}(k_i, k_j, m)\hat{F}_{;n}^{(m)}(\min(\hat{q}_{i;n}, \hat{q}_{j;n})) - \\ &- \sum_{m,l=1}^{M} w_{;n}(k_i, k_j, m, l)\hat{F}_{;n}^{(m)}(\hat{q}_{i;n})\hat{F}_{;n}^{(l)}(\hat{q}_{j;n}).\end{aligned} \tag{13}$$

Для оцінювання $f^{(m)}(q_i)$ можна скористатись ядерними оцінками щільності (п. 4.1 у [6]):

$$\hat{f}_{;n}^{(m)}(x; h) = \frac{1}{nh}\sum_{j=1}^{n} a_{j;n}^{(m)} K\left(\frac{x - \xi_{j;n}}{h}\right), \tag{14}$$

де K — ядро, тобто щільність деякого ймовірнісного розподілу, $h > 0$ — параметр згладжування. Відомо, що, якщо $f^{(m)}(x)$ є двічі неперервно диференційовною функцією x, то оптимальним ядром для її оцінювання є ядро Єпанєчнікова

$$K(x) = \begin{cases}(1 - x^2) & \text{при } |x| \le 1,\\ 0 & \text{при } |x| > 1.\end{cases}$$

Для вибору параметра згладжування h існує багато підходів. Одним із найбільш простих є правило Сілвермана. Ми розглянемо його модифікацію для оцінювання щільності за СЗК.

За теоремою 4.3.1 із [6], у випадку двічі неперервно диференційовних щільностей компонентів, асимптотично оптимальним (у розумінні мінімізації проінтегрованої середньоквадратичної похибки) є вибір параметра згладжування за формулою

$$h_{opt} = \left(\frac{A^{(m)}\|K\|^2}{nD^2(K)\phi(f^{(m)})}\right)^{1/5}. \tag{15}$$

Тут вважається, що

$$A^{(m)} = \lim_{n\to\infty} A_{;n}^{(m)},\ A_{;n}^{(m)} = \frac{1}{n}\sum_{j=1}^{n}(a_{j;n}^{(m)})^2,$$

$$\int_{-\infty}^{\infty} K(z)dz = 0,\ D(K) = \int_{-\infty}^{\infty} z^2 K(z)dz < \infty,\ \|K\|^2 = \int_{-\infty}^{\infty} (K(z))^2 dz < \infty,$$

$$\phi(f) = \int_{-\infty}^{\infty} \left(\frac{d^2}{dx^2} f(x)\right)^2 < \infty.$$

Використати безпосередньо h_{opt}, визначене (15), для оцінювання $f^{(m)}$ неможливо, оскільки для цього потрібно знати $\phi(f^{(m)})$. На практиці справжнє $\phi(f^{(m)})$ замінюють деякою оцінкою. Оскільки для щільності f гауссового розподілу з дисперсією σ^2,

$$\phi(f) = \frac{3}{8\sqrt{\pi}}\sigma^{-5},$$

то у звичайному правилі Сілвермана для оцінювання щільності за однорідною вибіркою використовується оцінка для $\phi(f)$:

$$\hat{\phi}(f) = \frac{3}{8\sqrt{\pi}} S^{-5},$$

де $S = S_0$, S_0 — вибіркове середньоквадратичне відхилення даних. Для того, щоб забезпечити робастність оцінки, використовують уточнене правило Сілвермана, в якому

$$S = S_1 = \min(S_0, \mathrm{IQR}/1.34),$$

де IQR — вибірковий інтерквартильний розмах даних.

У випадку, коли щільність оцінюється за вибіркою із суміші зі змінними концентраціями, природно замінити S_0 та IQR на відповідні оцінки середньоквадратичного відхилення та інтерквартильного розмаху:

$$\hat{S}^{(m)} = \sqrt{\frac{1}{n}\sum_{j=1}^{n} a_{j;n}^{(m)} (\xi_{j;n} - \bar{\xi}^{(m)})^2},\ \bar{\xi}^{(m)} = \frac{1}{n}\sum_{j=1}^{n} a_{j;n}^{(m)} \xi_{j;n},$$

$$\mathrm{IQR}^{(m)} = Q^{\hat{F}^{(m)}}(3/4) - Q^{\hat{F}^{(m)}}(1/4).$$

Остаточно, модифіковане правило Сілвермана для вибору параметра згладжування при оцінюванні щільності за СЗМ має такий вигляд:

$$\hat{h}_{Silv}^{(m)} = \left(\frac{8\sqrt{\pi} A_{;n}^{(m)} \|K\|^2}{3nD^2(K)}\right)^{1/5} \times \min\left(\hat{S}^{(m)}, \mathrm{IQR}^{(m)}\right). \tag{16}$$

Правило Сілвермана забезпечує близьку до оптимальної точність оцінювання якщо щільність, що оцінюється, не дуже сильно відрізняється від нормальної. Але навіть для щільностей дуже відмінних від нормальних, ядерні оцінки щільності з параметром згладжування, рівним $\hat{h}_{Silv}^{(m)}$, залишаються консистентними.

Тепер запишемо оцінку для s_{ij}:

$$\hat{s}_{ij;n} = \frac{\hat{v}_{ij;n}}{\hat{f}_{;n}^{(k_i)}(\hat{q}_{i;n}; \hat{h}_{Silv}^{(k_i)}) \hat{f}_{;n}^{(k_j)}(\hat{q}_{j;n}; \hat{h}_{Silv}^{(k_j)})},$$

$\hat{\mathbf{S}}_{;n} = (\hat{s}_{ij;n})_{ij=1}^{d}$ (тут вважається, що на роль ядра оцінки щільності вибрано ядро Єпанєчнікова).

Теорема 4.1. *При виконанні умов (**A**), (**B**), (**C**) матриця $\hat{\mathbf{S}}_{;n}$ є консистентною оцінкою $\mathbf{S}$.*

Доведення. Консистентність $\hat{v}_{ij;n}$ як оцінок v_{ij}, випливає з рівномірної (по x) консистентності $\hat{F}_{;n}^{(m)}(x)$ як оцінок $F^{(m)}(x)$ (наслідок 2.2.4 в [6]). Збіжність $\hat{f}_{;n}^{(k_i)}(\hat{q}_{i;n}; \hat{h}_{Silv}^{(k_i)}) \to f^{(k_i)}(q_i)$ випливає з консистентності $\hat{q}_{i;n}$ як оцінок (q_i) (за теоремою 3.1) і рівномірної консистентності ядерних оцінок щільності. □

Твердження про рівномірну консистентність ядерних оцінок щільності сформулюємо у вигляді окремої леми.

Лема 4.1. *Нехай виконуються такі умови.*

*1. Виконана умова (**A**).*

2. У т-го компонента суміші існує щільність розподілу $f^{(m)}$, неперервна на інтервалі $D_m = [q_m^-, q_m^+]$.

3. Ядро K є функцією обмеженої варіації і $K(x) = 0$ для всіх x таких, що $|x| > 1$.

4. $h_n = o(\sqrt{\log n/n})$ (за ймовірністю).

Тоді

$$\sup_{x\in D_m} |\hat f^{(m)}_{;n}(x; h_n) - f^{(m)}(x)| \to 0 \text{ при } n \to \infty.$$

Доведення. Позначимо

$$\bar f(x; h) = \int_{-\infty}^{\infty} \frac{1}{h} K\left(\frac{x-t}{h}\right) F^{(m)}(dt).$$

Позначимо як $V_{t\in\mathbb{R}}(K(t))$ варіацію функції K на R. Оцінимо

$$|\hat f^{(m)}_{;n}(x; h_n) - \bar f^{(m)}(x; h)| = \left|\int_{-\infty}^{\infty} \frac{1}{h} K\left(\frac{x-t}{h}\right)(\hat F^{(m)}_{;n}(dt) - F^{(m)}(dt))\right| \le$$
$$\le \sup_{t\in\mathbb{R}} |\hat F^{(m)}_{;n}(t) - F^{(m)}(t)| \times V_{t\in\mathbb{R}}\left(K\left(\frac{x-t}{h}\right)\right).$$

За наслідком 2.2.4 із [6], з урахуванням умови 1 леми, отримуємо

$$\sup_{t\in\mathbb{R}} |\hat F^{(m)}_{;n}(t) - F^{(m)}(t)| \le \Lambda\sqrt{\frac{\log n}{n}}, \tag{17}$$

де $\Lambda < \infty$ — деяка випадкова величина. Оскільки

$$V_{t\in\mathbb{R}}\left(K\left(\frac{x-t}{h}\right)\right) = \frac{V_{t\in\mathbb{R}}(K(t))}{h},$$

враховуючи умови 3 і 4 леми, маємо

$$\sup_{x\in\mathbb{R}} |\hat f^{(m)}_{;n}(x; h_n) - \bar f^{(m)}(x; h)| \to 0 \text{ за ймовірністю.} \tag{18}$$

Оцінимо тепер

$$\bar f^{(m)}(x; h) - f^{(m)}(x) = \int_{-1}^{1} K(z)(f(x - hz) - f(x))dz.$$

Враховуючи обмеженість K і рівномірну неперервність $f^{(m)}$ на D_m, отримуємо

$$\sup_{x\in D_m} |\bar f^{(m)}(x; h) - f^{(m)}(x)| \to 0 \text{ за ймовірністю при } h \to 0.$$

Звідси з урахуванням (18) випливає твердження леми. □

5. Довірчий еліпсоїд для вектора квантилів

Розглянемо задачу побудови довірчої множини для вектора $\mathbf{q}$, визначеного (9). Для цього скористаємось теоремами 3.1 і 4.1. Для $\mathbf{t} = (t_1, \dots, t_d)^T \in \mathbb{R}^d$ розглянемо функцію

$$R_{;n}(\mathbf{t}) = (\mathbf{t} - \hat{\mathbf{q}}_{;n})^T \hat{\mathbf{S}}^{-1}_{;n}(\mathbf{t} - \hat{\mathbf{q}}_{;n}).$$

Теорема 5.1. *Нехай виконані умови (**A**), (**B**), (**C**) і $\det \mathbf{S} \ne 0$. Тоді розподіл $R_{;n}(\mathbf{q})$ слабко збігається при $n \to \infty$ до розподілу χ^2 із d ступенями вільності.*

Доведення. Доведення безпосередньо випливає з теорем 3.1 і 4.1. □

Визначимо (асимптотичний) довірчий еліпсоїд рівня α для вектора невідомих параметрів $\mathbf{q}$ як множину

$$B_{;n}(\alpha) = \{\mathbf{t} \in \mathbb{R}^d \ : \ R_{;n}(\mathbf{t}) < Q^{\chi^2_d}(1-\alpha)\},$$

де $Q^{\chi^2_d}(1-\alpha)$ — квантиль розподілу χ^2 із d ступенями вільності рівня $1-\alpha$. Із теореми 5.1 випливає, що

$$\lim_{n\to\infty} \mathsf{P}\{\mathbf{q} \in B_{;n}(\alpha)\} \to 1-\alpha \text{ при } n \to \infty,$$

тобто $B_{;n}(\alpha)$ дійсно є довірчою множиною для $\mathbf{q}$ з асимптотичним рівнем значущості α.

6. Тести для перевірки гіпотез про квантилі

У загальному вигляді задачу перевірки гіпотез про вектор квантилів $\mathbf{q}$, визначений (9), можна сформулювати таким чином. Нехай задано деяку множину можливих значень $\mathbf{q}$, $Q_0 \subset \mathbb{R}^d$. Основна гіпотеза H_0 полягає в тому, що $\mathbf{q} \in Q_0$, альтернатива H_1 — $\mathbf{q} \in Q_1 = \mathbb{R}^d \setminus Q_0$.

Для перевірки таких гіпотез можна використати «графічний тест» π^{Graph} на основі довірчого еліпсоїда $B_{;n}(\alpha)$, побудованого у розділі 5. Тест приймає H_0, якщо $B_{;n}(\alpha) \cap Q_0 \neq \emptyset$, і відхиляє, якщо $B_{;n}(\alpha) \cap Q_0 = \emptyset$. Легко бачити, що асимптотична ймовірність помилки першого роду для цього тесту

$$\alpha_\infty(\pi^{Graph}) = \lim_{n\to\infty} \mathsf{P}_{\mathbf{q}}\{B_{;n}(\alpha) \cap Q_0 = \emptyset\} \leq \alpha \text{ при } \mathbf{q} \in Q_0.$$

Розглянемо тепер випадок "лінійних гіпотез", коли множину Q можна задати системою лінійних рівнянь. Нехай $l < d$ — деяке натуральне число, $\mathbf{L}$ — фіксована числова матриця з l рядків і d стовпців. Гіпотеза H_0 полягає в тому, що для вектора квантилів $\mathbf{q}$ виконується рівняння $\mathbf{Lq} = 0$, тобто

$$Q_0 = \{\mathbf{x} \in \mathbb{R}^d \ : \ \mathbf{Lx} = 0\}.$$

Окремими випадками цієї гіпотези є гіпотеза про рівність (однорідність) медіан різних компонентів суміші та гіпотеза про рівність інтерквартильних розмахів. Наприклад, у випадку трикомпонентної суміші ($M = 3$), гіпотезі про рівність інтерквартильних розмахів усіх трьох компонентів відповідають такі значення характеристик:

$$\mathbf{k} = (1,1,2,2,3,3), \beta = (3/4, 1/4, 3/4, 1/4, 3/4, 1/4),$$
$$\mathbf{L} = \begin{pmatrix} 1 & -1 & -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 & -1 & 1 \end{pmatrix}.$$

Таким чином, для лінійних гіпотез Q_0 є лінійним підпростором у $\mathbb{R}^d$. Зрозуміло, що для визначення Q_0 можна обмежитись матрицями $\mathbf{L}$ повного рангу. Надалі вважаємо, що ранг $\mathbf{L}$ дорівнює l.

Для лінійних гіпотез можна побудувати більш акуратний тест ніж π^{Graph}.

Визначимо норму (довжину) вектора $\mathbf{x}$ в $\mathbb{R}^d$ як

$$\|\mathbf{x}\|_{;n} = \sqrt{\mathbf{x}^T \hat{\mathbf{S}}^{-1}_{;n} \mathbf{x}}.$$

На роль статистики тесту виберемо квадрат відстані від оцінки вектора квантилів $\hat{\mathbf{q}}_{;n}$ до найближчого до неї елемента множини Q_0:

$$T_{;n} = n \inf_{\mathbf{x}:\ \mathbf{Lx}=0} \|\hat{\mathbf{q}}_{;n} - \mathbf{x}\|^2_{;n}.$$

Використовуючи метод множників Лагранжа, значення цієї статистики можна записати явно:

$$T_{;n} = n\hat{\mathbf{q}}^T_{;n} \mathbf{L}^T (\mathbf{L}\hat{\mathbf{S}}_{;n}\mathbf{L}^T)^{-1} \mathbf{L}\hat{\mathbf{q}}_{;n}.$$

Теорема 6.1. *Нехай виконані умови (**A**), (**B**), (**C**),* $\det \mathbf{S} \neq 0$ *і виконана гіпотеза* H_0*:* $\mathbf{Lq} = 0$*. Тоді розподіл* $T_{;n}$ *слабко збігається при* $n \to \infty$ *до розподілу* χ^2 *із* $d - l$ *ступенями вільності.*

Доведення. Позначимо

$$\langle \mathbf{x}, \mathbf{y} \rangle_{;n} = \mathbf{x}^T \hat{\mathbf{S}}_{;n}^{-1} \mathbf{y}$$

— скалярний добуток в $\mathbb{R}^d$, що відповідає нормі $\| \cdot \|_{;n}$,

$$Q_{;n}^{\perp} = \{\mathbf{y} \in \mathbb{R}^d \ : \ \langle \mathbf{x}, \mathbf{y} \rangle_{;n} = 0 \text{ для всіх } \mathbf{y} \in Q_0\}$$

— ортогональне доповнення Q_0 відносно скалярного добутку $\langle \cdot, \cdot \rangle_{;n}$. Тоді $T_{;n} = \| \sqrt{n} \Pr_{Q_{;n}^{\perp}} (\hat{\mathbf{q}}_{;n} - \mathbf{q}) \|_{;n}^2$, де $\Pr_A$ — оператор ортогонального проектування на підпростір A (тут ми скористались тим, що в умовах теореми $\mathbf{q} \in Q$, отже його проекція на $Q_{;n}^{\perp}$ дорівнює 0).

За теоремами 3.1, 4.1 розподіл $T_{;n}$ слабко збігається до розподілу $T_\infty = \| \Pr_{Q_\infty^{\perp}} \zeta \|_\infty^2$, де $\zeta \sim N(0, \mathbf{S})$, $\|\mathbf{x}\|_\infty^2 = \mathbf{x}^T \mathbf{S}^{-1} \mathbf{x}$,

$$\langle \mathbf{x}, \mathbf{y} \rangle_\infty = \mathbf{x}^T \mathbf{S}^{-1} \mathbf{y}, \ Q_\infty^{\perp} = \{\mathbf{y} \in \mathbb{R}^d \ : \ \langle \mathbf{x}, \mathbf{y} \rangle_\infty = 0 \text{ для всіх } \mathbf{y} \in Q_0\}.$$

Помітимо, що ζ є ізотропним гауссовим вектором з одиничною дисперсією у просторі $\mathbb{R}^2$ зі скалярним добутком $\langle \cdot, \cdot \rangle_\infty$ (тобто, для будь-якого вектора $\mathbf{x} \in \mathbb{R}^d$, $\langle \mathbf{x}, \zeta \rangle_\infty \sim N(0, \|\mathbf{x}\|^2)$). Тому $\Pr_{Q_\infty^{\perp}} \zeta$ є ізотропним гауссовим вектором в $Q_\infty^{\perp}$ і квадрат норми цього вектора має χ^2-розподіл із кількістю ступенів вільності, рівною вимірності простору $Q_\infty^{\perp}$.

Теорема доведена. □

Тепер можна визначити тест $\pi(T_{;n})$ для перевірки гіпотези H_0: $\mathbf{Lq} = 0$ проти альтернативи $\mathbf{Lq} \neq 0$. Для заданого рівня значущості α задамо поріг тесту $C_\alpha = Q^{\chi^2_{d-l}}(1 - \alpha)$. Тест приймає H_0, якщо $T_{;n} \leq C_\alpha$, і відхиляє, якщо $T_{;n} > C_\alpha$. За теоремою 6.1, асимптотичний рівень значущості цього тесту дорівнює α.

7. Результати імітаційного моделювання

Для перевірки того, наскільки акуратними є довірчі еліпсоїди і тести, запропоновані у розділах 5 і 6 при фіксованих обсягах вибірок, було проведено невелике експериментальне дослідження методом імітаційного моделювання.

Ми провели два експерименти. В обох розглядались суміші двох компонентів ($M = 2$). Концентрації першого компонента $p_{j;n}^1$ обирались як незалежні випадкові величини, рівномірно розподілені на $[0, 1]$. Концентрації другого компонента $p_{j;n}^2 = 1 - p_{j;n}^1$.

Для кожного обсягу вибірки, що розглядались, від $n = 500$ до $n = 10000$, було згенеровано по $B = 10000$ псевдовипадкових вибірок із суміші зі змінними концентраціями. На цих вибірках перевірялась точність асимптотичних еліпсоїдів і тестів.

Експеримент 1. У цьому експерименті перевіряється точність покриття довірчим еліпсоїдом справжнього значення невідомого параметра. Розподіли компонентів першого — $N(0, 2)$, другого — $N(1, 2)$. Оцінювались медіани першого і другого компонентів ($\mathbf{k} = (1, 2)$, $\beta = (1/2, 1/2)$). За кожною псевдовипадковою вибіркою було побудовано довірчий еліпсоїд $B_{;n}(\alpha)$ з $\alpha = 0,05$. Для заданих обсягів вибірки підраховувалась частота ν_n, з якою справжнє значення параметра $\mathbf{q} = (0, 1)$ потрапляло за межі цього еліпсоїда.

Результати експерименту вміщені у табл. 1.

Експеримент 2. Перевіримо, наскільки відрізняється частота помилок першого роду тесту $\pi(T_{;n})$, описаного у розділі 6 від номінального $\alpha = 0,05$. У цьому

експерименті $F^{(1)} \sim N(0,1)$, $F^{(1)} \sim N(0,2)$ перевіряється гіпотеза про рівність медіан $H_0\colon Q^{F^{(1)}}(1/2) = Q^{F^{(2)}}(1/2)$, тобто ($\mathbf{k} = (1,2)$, $\beta = (1/2, 1/2)$, $\mathbf{L} = (1,-1)$. За модельованими вибірками підраховувалась частота відхилення H_0 тестом $\pi(T_{;n})$. Результати експерименту представлені у табл. 2.

Таблиця 1. Частота виходу параметра з довірчого еліпсоїда

n	500	1000	2500	5000	10000
Частота	0,0262	0,0414	0,0412	0,0455	0,052

Таблиця 2. Частота помилок першого роду для тесту $\pi(T_{;n})$

n	500	1000	2500	5000	10000
Частота	0,0483	0,0384	0,0442	0,0462	0,0483

Як показують результати експериментів 1 і 2, точність асимптотичних формул невисока, але достатня для практичних потреб при обсягах вибірки більше 5000.

8. Висновки

Таким чином, ми побудували тести для перевірки гіпотез про квантилі розподілів компонентів у моделі суміші зі змінними концентраціями. Ці тести, зокрема, дозволяють перевіряти гіпотези про однорідність медіан або інтерквартильних розмахів компонентів. На жаль, результати імітаційного моделювання показують, що номінальний рівень значущості забезпечується у цих тестах лише при досить великих обсягах вибірки (не менше кількох тисяч). Тому актуальною є задача поліпшення якості таких тестів при помірних обсягах вибірки. Один із можливих напрямків для цього — застосування при оцінюванні асимптотичної дисперсії методів оцінювання щільності, що дають точність, кращу ніж ядерні оцінки з параметром згладжування, обраним за модифікованим правилом Сілвермана.

Список літератури

1. F. Autin, C. Pouet, *Minimax rates over Besov spaces in ill-conditioned mixture-models with varying mixing-weights*, Journal of Statistical Planning and Inference, **146** (2014), 20–30.
2. P. J. Bickel, M. J. Wichura, *Convergence criteria for multipapameter stochastic processes and some applications*, Ann. Math. Statist., **42** (1971), no. 5, 1656–1670.
3. A. Doronin, R. Maiboroda, *Testing hypotheses on moments by observations from a mixture with varying concentrations*, Modern Stochastics: Theory and Applications, **1** (2014), no. 2, 195–209.
4. M. Hollander, D. A. Wolfe, E. Chicken, *Nonparametric Statistical Methods*, John Wiley & Sons, Hoboken, 2014.
5. R. E. Maiboroda, *Statistical analysis of mixtures*, Kyiv University Publishers, Kyiv, 2003. (In Ukrainian)
6. R. E. Maiboroda, O. V. Sugakova, *Esitmation and classification by observations from a mixture*, Kyiv University Publishers, Kyiv, 2008. (In Ukrainian)
7. R. E. Maiboroda, *A test for the homogeneity of mixtures with varying concentrations*, Ukrainian Mathematical Journal, **52** (2000), no. 8, 1256–1263.
8. G. J. McLachlan, D. Peel, *Finite mixture models*, Wiley-Interscience, 2000.
9. G. J. McLachlan, K.-A. Do, C. Ambroise, *Analyzing Microarray Gene Expression Data*, Wiley-Interscience, 2004.
10. G. J. Myatt, W. P. Johnson, *Making Sense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining*, Willey, Hoboken, 2014.
11. A. Yu. Ryzhov, *A test of the hypothesis about the homogeneity of components of a mixture with varying concentrations by using censored data*, Theory Probab. Math. Statist., **72** (2006), 145–155.

12. J. Shao, *Mathematical statistics*, Springer-Verlag, New York, 1998.
13. J. W. Tukey, *Exploratory Data Analysis*, Addison Wesley, Reading MA, 1977.

Кафедра теорії ймовірностей, статистики та актуарної математики, механіко-математичний факультет, Київський національний університет імені Тараса Шевченка, вул. Володимирська, 64/13, м. Київ, Україна, 01601
Адреса електронної пошти: mre@univ.kiev.ua

Кафедра математики і теоретичної радіофізики, факультет радіофізики, електроніки і комп'ютерних систем, Київський національний університет імені Тараса Шевченка, вул. Володимирська, 64/13, м. Київ, Україна, 01601
Адреса електронної пошти: sugak@univ.kiev.ua

Стаття надійшла до редколегії 30.08.2019

TESTS ON QUANTILES OF MIXTURE COMPONENTS' DISTRIBUTIONS

R. E. MAIBORODA, O. V. SUGAKOVA

Abstract. The hypotheses testing problem is considered for hypotheses on quantiles of different components of a mixture with varying concentrations. Homogeneity of medians or interquartile ranges can be considered as the examples. Estimates for the quantiles are presented, their asymptotic normality is demonstrated. Asymptotic confidence ellipsoids and tests for linear hypotheses are constructed. Simulation results are presented.