Теорія ймовір. та матем. статист.
Вип. 86, 2012, стор. 182–192

Teor. Ĭmovir. ta Matem. Statyst.
No. 86, 2012, pp. 182–192

ОЦІНЮВАННЯ ПАРАМЕТРІВ БІНОМІАЛЬНОГО РОЗПОДІЛУ У МОДЕЛІ СУМІШІ

УДК 519.21

А. ЩЕРБІНА

Анотація. Розглядається модель спостережень з двокомпонентної суміші. З кожним об'єктом пов'язана певна числова характеристика. Она приймає значення нуль (невдача) або одиниця (успіх) з ймовірностями, спільними для всіх об'єктів одного класу. Спостерігаються декілька груп об'єктів. Відомі загальні кількості об'єктів першого та другого класів у групах та значення їх характеристик. Досліджується задача оцінювання ймовірностей успіху для обох компонентів. Для цього використовується метод максимальної вірогідності. Доведено консистентність та асимптотична нормальність відповідних оцінок. Для ілюстрації викладеної техніки оцінювання розглядається частковий випадок з галузі генетичних досліджень. Наводиться явний вигляд оцінки та асимптотична матриця розсіювання.

Abstract. Estimation of the mean value in the model of two-component mixture is considered. Each object has some some characteristic of interest which can be zero (fail) or one (success). Probabilities for this values are constant for objects from the same component. Objects are distributed over several groups. We know total quantities of objects of first and second components and their characteristics. Estimates of success probabilities for both components are constructed. Method of maximum likelihood is used. Consistency and asymptotic normality of estimates is proved.

Аннотация. Рассматривается модель наблюдений из двухкомпонентной смеси. С каждым объектом связана определенная числовая характеристика. Она принимает значение нуль (неудача) или единица (успех) с вероятностями, общими для всех объектов из одной компоненты. Наблюдается несколько групп объектов. Известны общие количества объектов первого и второго классов и значения их характеристик. Изучается задача оценивания вероятности успеха для обоих компонент. Для этого используется метод максимального правдоподобия. Доказана консистентность и асимптотическая нормальность оценок. Для иллюстрации изложенной техники оценивания рассматривается частный случай из сферы генетических исследований. Приводится явный вид оценки и асимптотической матрицы рассеивания.

1. Вступ

У даній роботі розглядається задача оцінювання параметрів моделі двокомпонентної суміші за спостереженнями, що розподілені по багатьом вибіркам (групам). При цьому кількість об'єктів, які належать певній компоненті суміші (класу) вважається відомою для всіх груп. Задачі такого роду природно виникають при аналізі даних соціологічних та медико-біологічних досліджень.

Як приклад, розглянемо дослідження залежності між успішністю учнів та їх ставленням до списування домашніх завдань. Оскільки питання про списування є "дражливим" (sensitive question), для визначення ставлення до нього проводиться анонімне опитування учнів по окремих групах (наприклад — по навчальних класах). В результаті такого опитування встановлюється кількість у кожній групі учнів, що допускають списування (перший компонент суміші) і тих, хто цього не допускає

2010 *Mathematics Subject Classification.* Primary 62F10, Secondary 62P10.

Ключові слова і фрази. Оцінювання у моделі суміші, параметричне оцінювання, генетичні дослідження.

(другий компонент). Крім того, для кожного учня за результатами навчання визначається рівень успішності, який, у найпростішому випадку, може бути бінарною характеристикою — 1 (успішний) або 0 (не успішний). Метою дослідження є оцінка середньої успішності (тобто ймовірності успіху) окремо для першого та для другого компонента.

Статистичний аналіз сумішей має довгу історію, починаючи з робіт Ньюкомба [9] та Пірсона [10]. Про сучасний стан проблеми див. книгу Маклахана та Піла [8]. Багатовибіркові задачі оцінювання характеристик сумішей розглядались у книзі Тітерінгтона та ін. [12]. Узагальненням цього підходу є модель суміші зі змінними концентраціями, розглянута у книзі Майбороди та Сугакової [4].

Особливістю розглядуваної нами задачі є те, що спостережувані об'єкти відбираються з невеликих скінченних популяцій (груп) без повернення, причому кількість об'єктів обох компонентів у цих групах є фіксованою та відомою. Тому між об'єктами існує залежність, і використання цього факту дозволяє більш точно оцінювати невідомі параметри моделі.

У статті розглядається оцінювання ймовірностей успіху для компонентів суміші за методом найбільшої вірогідності. Доведені консистентність та асимптотична нормальність отриманих оцінок. На відміну від звичайної техніки оцінювання за сумішами зі змінними концентраціями (що ігнорує залежність між об'єктами), запропонований підхід дає консистентні оцінки навіть тоді, коли концентрації компонентів у суміші є сталими для всіх груп об'єктів.

2. Постановка задачі

Нехай вибірка складається з K груп об'єктів обсягами $N_1, \dots, N_K$. Всі об'єкти розподілені за двома класами так, що в i-й групі знаходиться N_{i1} об'єктів першого класу та N_{i2} об'єктів другого класу. Загальну кількість об'єктів у i-й групі $N_{i1} + N_{i2}$ будемо позначати N_i. Пари (N_{i1}, N_{i2}) є незалежними однаково розподіленими векторами з розподілом

$$G(n_1, n_2) = \mathbf{P}\left(N_{i1} = n_1, N_{i2} = n_2\right), \qquad n_1, n_2 \in \mathbb{N}_0.$$

Звичайно, групи мають додатний розмір, тобто $G(0,0) = 0$. Також, будемо припускати наявність груп, що містять об'єкти обох класів одночасно. Тобто існують величини $n_1, n_2 > 0$ такі, що $G(n_1, n_2) > 0$.

Нехай C_{ij} — номер класу, до якого належить j-й об'єкт у i-тій групі. Вони не спостерігаються при обстеженні. При фіксованих N_{i1} та N_{i2} вектор $(C_{i1}, \dots, C_{iN_i})$ є таким, що всі його допустимі значення є рівноймовірними.

У кожного об'єкта спостерігається певна характеристика X, що може приймати значення нуль (невдача) або одиниця (успіх). Позначимо характеристики об'єктів у i-й групі $X_{i1}, \dots, X_{iN_i}$. Вони є випадковими величинами з розподілами, що залежать лише від класу, до якого належить об'єкт:

$$\mathbf{P}_q\left(X_i = 1 \mid C_i = m\right) = q_m \in [0,1], \qquad i = 1, \dots, N,\ m = 1, 2,$$

де q_m — ймовірність успіху для об'єкту з i-того класу. Позначимо $q = (q_1, q_2)$.

Задача полягає в оцінці невідомого параметра q за даними $\{N_{i1}, N_{i2}, X_{ij}, i = 1, \dots, K, j = 1, \dots, N_i\}$.

3. Дослідження випадкового механізму

Розглянемо i-ту групу. Позначимо суму характеристик об'єктів

$$X_i = \sum_{j=1}^{N_i} X_{ij}.$$

Покажемо, що статистика $S_i = (X_i, N_{i1}, N_{i2})$ є достатньою при оцінюванні параметра q за спостереженнями з i-той групи. Для цього досить довести, що ймовірність події $\{X_{ij} = x_j, j = 1, \dots, N_i\}$ залежить лише від суми $\sum_{j=1}^{N_i} x_i$. Для цього візьмемо два набори величин $x'_j, x''_j \in \{0, 1\}$, $j = 1, \dots, N_i$ з однаковими сумами:

$$\sum_{j=1}^{N_i} x'_j = \sum_{j=1}^{N_i} x''_j.$$

Оскільки кількості нулів та одиниць в обох наборах співпадають, то існує така перестановка σ, що $x''_j = x'_{\sigma(j)}$, $j = 1, \dots, N_i$. Тому виконуються наступні рівності:

$$\begin{aligned}\mathbf{P}_q\left(X_{ij} = x'_j, j = 1, \dots, N_i\right) &= \sum_{c_j \in \{1,2\}} \mathbf{P}_q\left(X_{ij} = x'_j, C_{ij} = c_j, j = 1, \dots, N_i\right) \\ &= \sum_{c_j \in \{1,2\}} \mathbf{P}_q\left(X_{ij} = x'_{\sigma(j)}, C_{ij} = c_{\sigma(j)}, j = 1, \dots, N_i\right) \\ &= \mathbf{P}_q\left(X_{ij} = x''_j, j = 1, \dots, N_i\right),\end{aligned}$$

де ми використали рівноймовірність всіх допустимих значень вектору $(C_{i1}, \dots, C_{iN_i})$.

Нехай значення невідомого параметра дорівнює $t = (t_1, t_2)$, а $s = (x, n_1, n_2)$. Введемо наступні позначення:

$$f(s,t) = f(x, n_1, n_2, t) = \mathbf{P}_t\left(X_i = x, N_{i1} = n_1, N_{i2} = n_2\right),$$
$$g(s,t) = g(x, n_1, n_2, t) = \mathbf{P}_t\left(X_i = x \mid N_{i1} = n_1, N_{i2} = n_2\right).$$

Позначивши $n = n_1 + n_2$, обчислимо ці ймовірності:

$$\begin{aligned}g(s,t) &= \sum_{k=0}^{x} \mathbf{P}_t\left(\sum_{j=1}^{n} X_{ij}\, \mathbf{1}_{\{C_{ij}=1\}} = k\right) \mathbf{P}_t\left(\sum_{j=1}^{n} X_{ij}\, \mathbf{1}_{\{C_{ij}=2\}} = x - k\right) \\ &= \sum_{k=0}^{x} C_{n_1}^{k} t_1^k (1-t_1)^{n_1-k} C_{n_2}^{x-k} t_2^{x-k} (1-t_2)^{n_2-x+k}\, \mathbf{1}_{\{x-n_2 \le k \le n_1\}}.\end{aligned}$$

Також

$$f(s,t) = \mathbf{P}_t\left(X = x \mid N_1 = n_1, N_2 = n_2\right) G(n_1, n_2) = g(s,t) G(n_1, n_2).$$

4. Оцінка максимальної вірогідності

Нехай $S = (S_1, \dots, S_K)$. Тоді логарифмічна функція вірогідності дорівнюватиме

$$L(S,t) = \sum_{i=1}^{K} \ln f(S_i, t) = \sum_{i=1}^{K} l(S_i, t), \tag{1}$$

де $l(S_i, t) = \ln f(S_i, t)$.

Отже, оцінкою методу максимальної вірогідності буде значення параметра t, що максимізує вираз (1):

$$\hat{q} = \operatorname*{argmax}_{t \in [0,1]^2} L(S,t).$$

5. Консистентність

Нехай q — справжнє значення невідомого параметра. При дослідженні асимптотичної поведінки оцінки максимальної вірогідності ми маємо розрізняти два змістовно різних випадки:

1. Розміри класів співпадають у всіх групах $N_{11} = N_{12}$ м.н.
2. Є групи з різними розмірами класів $\mathbf{P}_q\left(N_{11} \neq N_{12}\right) > 0$.

В першому випадку задача перестає бути ідентифікованою, адже функція вірогідності стає симетричною:

$$L(S, (t_1, t_2)) = L(S, (t_2, t_1)).$$

Отже, розв'язок можна визначити лише з точністю до перестановки. Нехай для визначеності $q_1 < q_2$. В такому випадку максимум функції вірогідності будемо шукати серед тих $t \in [0,1]^2$, для яких $t_1 \le t_2$:

$$\hat{q}^* = \operatorname*{argmax}_{0 \le t_1 \le t_2 \le 1} L(S,t).$$

У другому ж випадку зрештою розв'язок стає єдиним. Виконується наступна

Теорема 1. *Нехай обсяги груп мають скінченні математичні сподівання*

$$\mathbf{E}_q N_1 < \infty.$$

Якщо $N_{11} = N_{12}$ м.н., то оцінка максимальної вірогідності $\hat{q}^*$ є строго консистентною. Якщо $\mathbf{P}_q(N_{11} \ne N_{12}) > 0$, то оцінка методу максимальної вірогідності $\hat{q}$ є строго консистентною.

При доведенні цієї теореми ми будемо розглядати відстань Кульбака–Лейблера між розподілами $f(s,t)$ та $f(s,q)$:

$$\rho(t) = \sum_s f(s,q) \ln \frac{f(s,q)}{f(s,t)} = \mathbf{E}_q l(S_1,q) - \mathbf{E}_q l(S_1,t)$$

Скористаємось наступним твердженням.

Твердження 1 ([1, Теорема 16.2])**.** *Нехай виконуються умови*

(1) *Параметрична множина Θ компактна,*
(2) *Функція $\rho(t)$ досягає свого мінімуму в єдиній точці $t = q$,*
(3) *Функція $f(s,t)$ є диференційованою за t, та для всіх $t \in \Theta$ виконується*

$$\sum_s \ln \frac{f^\Theta(s)}{f(s,q)} f(s,q) < \infty, \qquad f^\Theta(s) = \sup_{t \in \Theta} f(s,t).$$

Тоді оцінка максимальної вірогідності $\hat{q}$ строго консистентна.

Отже, для доведення консистентності ми маємо дослідити поведінку функції ρ на мінімум. Для цього наведемо наступну лему.

Лема 1 ([1, Лема 16.1])**.** *Нехай f та g — дві щільності ймовірності відносно міри μ. Тоді*

$$\int f(x) \ln f(x)\, \mu(dx) \ge \int f(x) \ln g(x)\, \mu(dx),$$

якщо обидва ці інтеграли скінченні. Знак рівності можливий лише у випадку $f = g$ м.с. за мірою μ.

В нашому випадку міра μ — рахуюча на множині (x, n_1, n_2) для $n_1, n_2 \ge 0$ та $0 \le x \le n_1 + n_2$. Отже, значення функції ρ мінімальне в точці q. Якщо ж значення параметра t також мінімізує функцію ρ, то для всіх $n_1, n_2 \ge 0$ та $0 \le x \le n_1 + n_2$ має виконуватися

$$f(x, n_1, n_2, t) = f(x, n_1, n_2, q).$$

При цьому ця рівність буде змістовною лише коли $G(n_1, n_2) > 0$. Для таких обсягів n_1 та n_2 її можна переписати наступним чином:

$$g(x, n_1, n_2, t) = g(x, n_1, n_2, q). \tag{2}$$

При доведенні Теореми 1 нам знадобляться такі леми.

Лема 2. *Нехай $N_{11} = N_{12}$ м.н. Тоді функція ρ досягає свого мінімуму лише в точках (q_1, q_2) та (q_2, q_1).*

Лема 3. *Нехай $\mathbf{P}_q(N_{11} = N_{12}) < 1$ м.н. Тоді функція ρ досягає свого мінімуму в єдиній точці (q_1, q_2).*

Доведення Теореми 1. Якщо $N_{11} = N_{12}$ м.н., то візьмемо $\Theta = \{(q_1, q_2) \in [0,1]^2 \mid q_1 \le q_2\}$. За Лемою 2 функція $\rho(t)$ досягає свого мінімуму в точках (q_1, q_2) та (q_2, q_1). А оскільки $q_1 \le q_2$, то мінімум функції $\rho(t)$ на множині Θ єдиний і дорівнює q.

Якщо ж $\mathbf{P}_q(N_{11} = N_{12}) < 1$, то візьмемо $\Theta = [0,1]^2$. За Лемою 3 функція $\rho(t)$ досягає свого мінімуму в єдиній точці q.

Отже, умови (1) та (2) Твердження 1 виконуються. Очевидно, що функції $f(s,t)$ є диференційованими за t для всіх s. Використовуючи нерівність $x \ln x > -1$, отримуємо:

$$\sum_s \ln \frac{f^\Theta(s)}{f(s,q)} f(s,q) = \sum_s \ln \frac{\sup_{t\in\Theta} g(s,t)}{g(s,q)} f(s,q) \le -\sum_s f(s,q) \ln g(s,q)$$

$$= -\sum_{n_1,n_2\ge 0} G(n_1,n_2) \sum_{x=0}^{n_1+n_2} g(s,q) \ln g(s,q) \le \sum_{n_1,n_2\ge 0} G(n_1,n_2) n = \mathbf{E}_q N_1.$$

Отже, виконуються всі умови Твердження 1, і оцінка максимальної вірогідності є строго консистентною. □

6. Асимптотична нормальність

Теорема 2. *Нехай обсяги груп мають скінченний другий момент $\mathbf{E}_q N_1^2 < \infty$, а справжнє значення $q \in (0,1)^2$. Тоді маємо такі випадки:*

(1) *Розміри класів співпадають у всіх групах $N_{11} = N_{12}$ м.н. Якщо $q_1 < q_2$, то оцінка максимальної вірогідності $\hat{q}^*$ є асимптотично нормальною.*
(2) *Є групи з різними розмірами класів $\mathbf{P}_q(N_{11} \ne N_{12}) > 0$. Якщо $q_1 \ne q_2$ або не існує числа $C > 0$ такого, що $N_{11} = CN_{12}$ м.н., то оцінка методу максимальної вірогідності $\hat{q}$ є асимптотично нормальною.*

Для доведення використаємо наступне твердження:

Твердження 2 ([1, Теорема 14.4])**.** *Нехай виконані наступні умови:*

(1) *Параметрична множина Θ компактна,*
(2) *Функція $l(s,t)$ двічі неперервно диференційована за t,*
$$\sup_{t\in\Theta} \left| l''_{t_i t_j}(s,t) \right| < \gamma(s), \qquad i,j = 1,2,\ \mathbf{E}_q \gamma(S_1) < \infty.$$
(3) *Існує матриця $I = \mathbf{E}_q \frac{\partial^2}{\partial t^2} l(S_1, q)$, $\det I \ne 0$.*
(4) *Рівняння $\mathbf{E}_q \frac{\partial}{\partial t} l(S_1, t) = 0$ має єдиний розв'язок $t = q$.*

Тоді оцінка максимальної вірогідності $\hat{q}$ є асимптотично нормальною:
$$\sqrt{K}(\hat{q}_K - q) \to \mathcal{N}(0, I^{-1}).$$

Доведення Теореми 2. Візьмемо параметричну множину Θ так, як в доведенні Теореми 1. Тоді оцінка максимальної вірогідності є консистентною, тому умови Твердження 2 достатньо перевірити в деякому околі істинного значення q.

В нашому випадку функція $l(s,t)$ є двічі неперервно диференційованою. Зафіксуємо довільне $\varepsilon > 0$ таке, що $q \in (\varepsilon, 1-\varepsilon)^2$. Покажемо виконання умови (2) Твердження 2 для $t \in [\varepsilon, 1-\varepsilon]^2$. Запишемо,

$$l''_{t_i t_j}(s,t) = \frac{f''_{t_i t_j}(s,t)}{f(s,t)} - \frac{f'_{t_i}(s,t)}{f(s,t)} \frac{f'_{t_j}(s,t)}{f(s,t)} = \frac{g''_{t_i t_j}(s,t)}{g(s,t)} - \frac{f'_{t_i}(s,t)}{g(s,t)} \frac{f'_{t_j}(s,t)}{g(s,t)}.$$

Оцінимо частки в правій частині:

$$\left| \frac{g'_{t_1}(s,t)}{g(s,t)} \right| \le \frac{\sum_{k=0}^x \left| \frac{k}{t_1} - \frac{n_1-k}{1-t_1} \right| C_{n_1}^k t_1^k (1-t_1)^{n_1-k} C_{n_2}^{x-k} t_2^{x-k} (1-t_2)^{n_2-x+k} \mathbf{1}_{\{x-n_2\le k\le n_1\}}}{\sum_{k=0}^x C_{n_1}^k t_1^k (1-t_1)^{n_1-k} C_{n_2}^{x-k} t_2^{x-k} (1-t_2)^{n_2-x+k} \mathbf{1}_{\{x-n_2\le k\le n_1\}}}$$

$$\le \frac{2n}{\varepsilon}.$$

Аналогічно можна отримати наступні нерівності:

$$\left|\frac{g'_{t_2}(s,t)}{g(s,t)}\right| \le \frac{2n}{\varepsilon}, \qquad \left|\frac{g''_{t_it_j}(s,t)}{g(s,t)}\right| \le \frac{4n^2}{\varepsilon^2}, \quad i,j=1,2.$$

Отже, маємо:

$$\left|l''_{t_it_j}(s,t)\right| \le \frac{4n^2}{\varepsilon^2} + \frac{2n}{\varepsilon}\frac{2n}{\varepsilon} = \frac{8n^2}{\varepsilon^2}.$$

Зі скінченності другого моменту від обсягу групи випливає виконання умови (2).

Розглянемо матрицю $I = \mathbf{E}_q \frac{\partial^2}{\partial t^2} l(S_1, q)$. Покажемо, що в умовах теореми вона є невиродженою. Розпишемо елемент (i,j) матриці I:

$$\begin{aligned} I_{ij} &= \mathbf{E}_q \left[\frac{f''_{t_it_j}(S_1,q)}{f(S_1,q)} - \frac{f'_{t_i}(S_1,q) f'_{t_j}(S_1,q)}{f^2(S_1,q)} \right] = \sum_s \left[f''_{t_it_j}(s,q) - \frac{f'_{t_i}(s,q) f'_{t_j}(s,q)}{f(s,q)} \right] \\ &= \frac{\partial^2}{\partial t_i \partial t_j} \sum_s f(s,q) - \sum_s \frac{f'_{t_i}(s,q) f'_{t_j}(s,q)}{f(s,q)} = -\sum_s \frac{f'_{t_i}(s,q) f'_{t_j}(s,q)}{f(s,q)}. \end{aligned} \tag{3}$$

Визначник матриці I дорівнює $I_{11}I_{22} - I_{12}^2$ і завжди невід'ємний, адже за нерівністю Коші–Буняковського виконується

$$\left(\sum_s \frac{g'_{t_1}(s,q) g'_{t_2}(s,q)}{g(s,q)} \right)^2 \le \left(\sum_s \frac{g'_{t_1}(s,q) g'_{t_1}(s,q)}{g(s,q)} \right) \left(\sum_s \frac{g'_{t_2}(s,q) g'_{t_2}(s,q)}{g(s,q)} \right).$$

При цьому рівність досягається лише коли функції $\frac{\partial}{\partial t_1} g(s,q)$ та $\frac{\partial}{\partial t_2} g(s,q)$ пропорційні. Покажемо, що в умовах теореми ці функції пропорційними не будуть.

Візьмемо довільні $n_1, n_2 > 0$ такі, що $G(n_1, n_2) > 0$. Розглянемо відношення функцій $\frac{\partial}{\partial t_1} g(s,q)$ та $\frac{\partial}{\partial t_2} g(s,q)$ для $s' = (0, n_1, n_2)$ та $s'' = (n_1 + n_2, n_1, n_2)$:

$$\frac{g'_{t_1}(s',q)}{g'_{t_2}(s',q)} = \frac{-n_1(1-q_1)^{n_1-1}(1-q_2)^{n_2}}{-n_2(1-q_1)^{n_1}(1-q_2)^{n_2-1}} = \frac{n_1(1-q_2)}{n_2(1-q_1)},$$

$$\frac{g'_{t_1}(s'',q)}{g'_{t_2}(s'',q)} = \frac{-n_1 q_1^{n_1-1} q_2^{n_2}}{-n_2 q_1^{n_1} q_2^{n_2-1}} = \frac{n_1 q_2}{n_2 q_1}$$

Якщо $q_1 \ne q_2$, то для цих часток ми вже отримаємо порушення пропорційності. Якщо ж $q_1 = q_2$, то можна переконатися, що функції $\frac{\partial}{\partial t_1} g(s,q)$ та $\frac{\partial}{\partial t_2} g(s,q)$ є пропорційними тоді і лише тоді, коли існує число $C > 0$ таке, що $N_1 = CN_2$ м.н.

Розглянемо відображення $s(t) = \mathbf{E}_q \frac{\partial}{\partial t} l(S_1, t)$. В точці q воно дорівнює нулю:

$$s(q) = \mathbf{E}_q \frac{\partial}{\partial t} l(S_1, q) = \sum_s f(s,q) \frac{\frac{\partial}{\partial t} f(s,q)}{f(s,q)} = \frac{\partial}{\partial t} \sum_s f(s,q) = \frac{\partial}{\partial t} 1 = 0.$$

З невиродженості матриці I випливає, що в деякому околі точки q розв'язок рівняння $s(t) = 0$ єдиний.

Для перевірки умови (5) запишемо:

$$\mathbf{Cov}_q \frac{\partial}{\partial t} l(S_1, q) = \mathbf{E}_q \frac{\partial}{\partial t} l(S_1, q) \left(\frac{\partial}{\partial t} l(S_1, q) \right)^T - s(q) s(q)^T = -I.$$

Отже, всі умови Теореми 2 виконуються, тому оцінка методу максимальної вірогідності є асимптотично нормальною. □

7. Приклад застосування

Розглянемо частковий випадок, коли кожна група складається з двох об'єктів, по одному з першого та другого класів, тобто $G(1,1) = 1$.

Така модель може бути корисною, наприклад, при дослідженні явища геномного імпринтингу. Воно полягає в тому, що наявність або відсутність експресії дієздатного гену в організмі у деяких випадках залежить від того, на якій хромосомі розміщена дана копія (алель) гену: на успадкованій від батька, чи від матері.

Нехай досліджується експресія певного гену у гомозиготних організмах, причому за фенотипом можна встановити, чи має місце у даному (i-тому) організмі експресія генів з однієї ($X_i = 1$) обох ($X_i = 2$) або жодної ($X_i = 0$) з хромосом. У цьому випадку кожен організм можна розглядати як окрему групу, що містить два об'єкти — хромосоми, одна з яких належить до першого компонента (успадковані від батька), а друга — до другого (материнська). Значення X_{ij} вважаємо рівним 1, якщо у i-тому організмі має місце експресія гену на j-тій хромосомі і 0 в іншому випадку. Тоді $X_i = X_{i1} + X_{i2}$. Хоча ми не спостерігаємо X_{ij} окремо, але, як показано вище, X_i є достатньою статистикою для побудови оцінок імовірностей q_k того, що матиме місце експресія гену з k-того компонента.

Дослідимо цю задачу за допомогою методу максимальної вірогідності. Введемо наступні позначення:

$$f_i(t) = f(i, 1, 1, t) = g(i, 1, 1, t), \qquad i = 0, 1, 2.$$

Тоді функцію вірогідності (1) можна переписати у наступному вигляді:

$$L(S, t) = \sum_{i=1}^{K} \ln f_{X_i}(t) = K \sum_{l=0}^{2} \nu_l \ln f_l(t),$$

де ν_l — це частота значень l у вибірці $X_1, \ldots, X_K$.

Оскільки ν_l та $f_l(t)$ є розподілами ймовірностей на множині $\{0, 1, 2\}$, то за Лемою 1 максимум досягається коли $\nu_l = f_l(t)$. Для $l = 1, 2$ отримаємо наступні рівняння:

$$\nu_1 = f_1(t) = t_1(1 - t_2) + (1 - t_1)t_2, \tag{4}$$

$$\nu_2 = f_2(t) = t_1 t_2. \tag{5}$$

Виражаючи t_2 з другого рівняння та підставляючи в друге, отримаємо таке квадратне рівняння:

$$t_1^2 - (\nu_1 + 2\nu_2)t_1 + \nu_2 = 0.$$

При $\nu_2 \le \mu^2/4$ це рівняння має розв'язки $t = \mu/2 \pm \sqrt{\mu^2/4 - \nu_2}$, де μ — це середнє значення за вибіркою:

$$\mu = \frac{1}{K} \sum_{i=1}^{K} x_i = \nu_1 + 2\nu_2.$$

Тому отримаємо наступну оцінку:

$$\hat{q}^* = \left(\frac{\mu}{2} - \sqrt{\frac{\mu^2}{4} - \nu_2}, \frac{\mu}{2} + \sqrt{\frac{\mu^2}{4} - \nu_2} \right),$$

Введемо в розгляд наступні множини:

$$A = \left\{ (\nu_1, \nu_2) \mid 0 \le \nu_1, \nu_2 \le 1, \nu_1 + \nu_2 \le 1, \nu_2 \le (\nu_1 + 2\nu_2)^2/4 \right\},$$

$$B = \left\{ (\nu_1, \nu_2) \mid 0 \le \nu_1, \nu_2 \le 1, \nu_1 + \nu_2 \le 1, \nu_2 > (\nu_1 + 2\nu_2)^2/4 \right\}.$$

Вони зображені на Рис. 1. Оскільки рівняння (4)–(5) мають розв'язок тоді і тільки тоді, коли $(\nu_1, \nu_2) \in A$, то для множини A можна записати

$$A = \left\{ (f_1(t), f_2(t)) \mid t \in [0, 1]^2 \right\}. \tag{6}$$

Тепер розглянемо випадок $(\nu_1, \nu_2) \in B$. Перепишемо функцію вірогідності (1) як функцію від ν_1, ν_2, $f_1(t)$ та $f_2(t)$:

$$L(S, t) = K \left[\nu_0 (1 - f_1(t) - f_2(t)) + \nu_1 \ln f_1(t) + \nu_2 \ln f_2(t) \right].$$

Використовуючи рівність (6), задачу максимізації функції вірогідності можна переписати наступним чином:

$$\begin{cases} F(p_1, p_2) := \nu_0 \ln(1 - p_1 - p_2) + \nu_1 \ln p_1 + \nu_2 \ln p_2 \to \max, \\ (p_1, p_2) \in A. \end{cases}$$

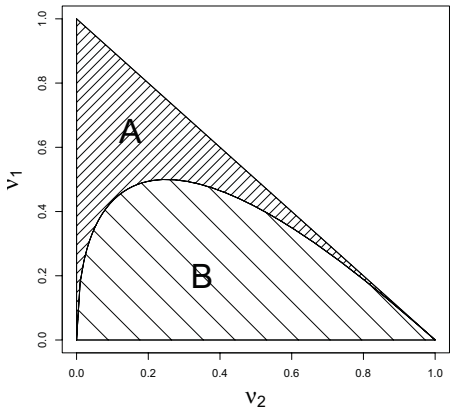

Рис. 1. Множина значень вектора (ν_1, ν_2) для яких рівняння (4-5) мають розв'язок

Покажемо, що функція $F(p_1, p_2)$ є угнутою за p_1, p_2. Її матриця других похідних має вигляд:

$$\frac{\partial^2 F(p_1,p_2)}{\partial p_1 \partial p_2} = \begin{pmatrix} -\dfrac{\nu_0}{(1-p_1-p_2)^2} - \dfrac{\nu_1}{p_1^2} & -\dfrac{\nu_0}{(1-p_1-p_2)^2} \\ -\dfrac{\nu_0}{(1-p_1-p_2)^2} & -\dfrac{\nu_0}{(1-p_1-p_2)^2} - \dfrac{\nu_2}{p_2^2} \end{pmatrix}.$$

Оскільки перший мінор від'ємний, а визначник додатний, то за критерієм Сильвестра отримуємо, що вона є від'ємно напіввизначеною. Тому для $(\nu_1, \nu_2) \in B$ максимум функції $F(p_1, p_2)$ буде досягатися на криволінійній межі множини A. Можна переконатися, що будь-якої точки цієї межі виконується $(p_1, p_2) = (f_1(s,s), f_2(s,s))$ для деякого $s \in [0,1]$. Отже, функція $L(S,t)$ досягає свого максимуму при $t_1 = t_2$. Тому ми приходимо до наступної задачі:

$$\begin{cases} G(s) = \nu_0 \ln(1-s)^2 + \nu_1 \ln 2s(1-s) + \nu_2 \ln s^2 \to \max, \\ s \in [0,1]. \end{cases}$$

Похідна функції $G(s)$ дорівнює:

$$G'(s) = -\frac{2\nu_0}{1-s} + \frac{\nu_1}{s} - \frac{\nu_1}{1-s} + \frac{2\nu_2}{s}.$$

Можна пересвідчитись, що єдиним коренем рівняння $G'(s) = 0$ буде $s = \mu/2$. Тому оцінка максимальної вірогідності у випадку $\nu_2 > \mu^2/4$ має вигляд $\hat{q}^* = (\mu/2, \mu/2)$.

З Теореми 1 випливає, що отримана оцінка $\hat{q}^*$ є консистентною при $q_1 \le q_2$. А З Теореми 2 випливає асимптотична нормальність при $0 < q_1 < q_2 < 1$. При цьому за формулою (3) інформаційна матриця дорівнює:

$$I = \frac{1}{q_1 + q_2 - 2q_1q_2} \begin{pmatrix} \dfrac{q_1 + q_2^2 - 2q_1q_2}{q_1 - q_1^2} & 1 \\ 1 & \dfrac{q_1^2 + q_2 - 2q_1q_2}{q_2 - q_2^2} \end{pmatrix}.$$

8. Висновки

Досліджено задачу оцінювання у моделі двокомпонентної суміші. Побудовано і досліджено оцінки методу максимальної вірогідності. Наведено явний вигляд оцінки та асимптотична матриця розсіювання у частковому випадку.

Подальшим напрямком досліджень є застосування розвиненої техніки до сумішей з довільним розподілом компонентів. Якщо $\{X_{ij}, j = 1, \dots, N_i\}$ — спостережувані дані з довільним розподілом, то можна перейти до набору

$$\{Y_{ij} = \mathbf{1}_{\{X_{ij}<x\}}, j = 1, \dots, N_i\}$$

для довільного x. До цих даних можна застосувати метод максимальної вірогідності. Отримані оцінки є значеннями функцій розподілу компонентів в точці x. Таким чином можна оцінити розподіл компонентів суміші, а отже й будь-які ймовірнісні характеристики компонентів.

9. Доведення лем

Доведення леми 2. Неважко переконатися, що для всіх $n_1 > 0$ та $0 \le x \le 2n_1$ виконується

$$g(x, n_1, n_1, (q_1, q_2)) = g(x, n_1, n_1, (q_2, q_1)).$$

Отже, набір (q_2, q_1) також мінімізує функцію ρ. Покажемо, що функція ρ досягає свого мінімуму лише в цих точках.

Розглянемо випадок $q \in (0,1)^2$. Припустимо, що функція ρ досягає свого мінімуму у точці $t = (t_1, t_2)$. З умови леми випливає, що існує таке число n, що $G(n, n) > 0$. Застосуємо умову (2) до наборів $(0, n, n)$ та $(2n, n, n)$:

$$(1-t_1)^n(1-t_2)^n = (1-q_1)^n(1-q_2)^n,$$
$$t_1^n t_2^n = q_1^n q_2^n.$$

Або, скорочуючи степені та розкриваючи дужки:

$$1 - t_1 - t_2 + t_1 t_2 = 1 - q_1 - q_2 + q_1 q_2$$
$$t_1 t_2 = q_1 q_2.$$

Звідки маємо:

$$t_1 + t_2 = q_1 + q_2, \qquad t_1 t_2 = q_1 q_2.$$

За теоремою Вієта ця система має лише два розв'язки (q_1, q_2) та (q_2, q_1).

Тепер розглянемо випадок, коли параметр q лежить на межі квадрату $[0,1]^2$. Нехай, наприклад, $q_1 = 0$. Тоді рівності (2) для наборів $(0, n, n)$ та $(2n, n, n)$ матимуть вигляд:

$$(1-t_1)^n(1-t_2)^n = (1-q_2)^n,$$
$$t_1^n t_2^n = 0.$$

Звідси знаходимо два розв'язки $(0, q_2)$ та $(q_2, 0)$. У інших випадках доведення аналогічне. □

Доведення леми 3. Припустимо, що функція ρ також досягає свого мінімуму у точці $t = (t_1, t_2)$. Розглянемо випадок $q \in (0,1)^2$.

Запишемо умову (2) для n_1 та n_2 з умови теореми та $x = 0,\ 1,\ n_1 + n_2 - 1,\ n_1 + n_2$:

$$(1-t_1)^{n_1}(1-t_2)^{n_2} = (1-q_1)^{n_1}(1-q_2)^{n_2}, \tag{7}$$

$$\begin{aligned} n_1 t_1 (1-t_1)^{n_1-1}(1-t_2)^{n_2} + n_2(1-t_1)^{n_1} t_2 (1-t_2)^{n_2-1} \\ = n_1 q_1 (1-q_1)^{n_1-1}(1-q_2)^{n_2} + n_2(1-q_1)^{n_1} q_2 (1-q_2)^{n_2-1}, \end{aligned} \tag{8}$$

$$\begin{aligned} n_1 t_1^{n_1-1}(1-t_1)t_2^{n_2} + n_2 t_1^{n_1} t_2^{n_2-1}(1-t_2) \\ = n_1 q_1^{n_1-1}(1-q_1)q_2^{n_2} + n_2 q_1^{n_1} q_2^{n_2-1}(1-q_2), \end{aligned} \tag{9}$$

$$t_1^{n_1} t_2^{n_2} = q_1^{n_1} q_2^{n_2}. \tag{10}$$

Розділивши (8) на (7), та (9) на (10), отримаємо:

$$n_1\frac{t_1}{1-t_1}+n_2\frac{t_2}{1-t_2}=n_1\frac{q_1}{1-q_1}+n_2\frac{q_2}{1-q_2},$$
$$n_1\frac{1-t_1}{t_1}+n_2\frac{1-t_2}{t_2}=n_1\frac{1-q_1}{q_1}+n_2\frac{1-q_2}{q_2}.$$

Введемо позначення

$$\alpha=\frac{n_1}{n_2},\qquad p_i=\frac{1-t_i}{t_i},\quad r_i=\frac{1-q_i}{q_i},\quad i=1,2.$$

Маємо $\alpha\neq 1$ та

$$\frac{\alpha}{p_1}+\frac{1}{p_2}=\frac{\alpha}{r_1}+\frac{1}{r_2},$$
$$\alpha p_1+p_2=\alpha r_1+r_2. \tag{11}$$

Перемноживши ці рівності між собою, отримаємо:

$$\alpha^2+\alpha\left(\frac{p_1}{p_2}+\frac{p_2}{p_1}\right)+1=\alpha^2+\alpha\left(\frac{r_1}{r_2}+\frac{r_2}{r_1}\right)+1,$$

або

$$\frac{p_1}{p_2}+\frac{p_2}{p_1}=\frac{r_1}{r_2}+\frac{r_2}{r_1}.$$

Якщо $r_1=r_2$, то отримаємо єдиний розв'язок $p_1=p_2$. З (11) випливає, що $p_1=p_2=r_1$. Отже, функція ρ має єдину точку максимуму $t=q$.

Нехай $r_1\neq r_2$. Неважко показати, що відносно p_1/p_2 це рівняння має два корені, і вони дорівнюють r_1/r_2 та r_2/r_1. Тому маємо $p_2=p_1r_2/r_1$ або $p_2=p_1r_1/r_2$. Розділивши (7) на (10), отримаємо:

$$p_1^{\alpha}p_2=r_1^{\alpha}r_2.$$

Підставивши в це рівняння значення для p_2, отримаємо $p_1=r_1$ або $p_1=r_1^{\frac{\alpha-1}{\alpha+1}}r_2^{\frac{2}{\alpha+1}}$. У першому випадку маємо $p_2=r_2$, а тому $t=q$. У другому випадку $p_2=r_1^{\frac{2\alpha}{\alpha+1}}r_2^{\frac{1-\alpha}{\alpha+1}}$. Покажемо, що це є хибний розв'язок. Підставимо його у перше рівняння (11):

$$\alpha r_1^{\frac{\alpha-1}{\alpha+1}}r_2^{\frac{2}{\alpha+1}}+r_1^{\frac{2\alpha}{\alpha+1}}r_2^{\frac{1-\alpha}{\alpha+1}}=\alpha r_1+r_2.$$

Позначимо $v=(r_1/r_2)^{1/(\alpha+1)}$, маємо $v\neq 1$. Розділивши попередню рівність на r_2 отримаємо:

$$\alpha v^{\alpha-1}+v^{2\alpha}=\alpha v^{\alpha+1}+1,$$

або, розділивши на αv^{α},

$$\frac{v^{\alpha}-v^{-\alpha}}{\alpha}=v-v^{-1}. \tag{12}$$

Розглянемо функцію $F(v,\alpha)=\left(v^{\alpha}-v^{-\alpha}\right)/\alpha-(v-v^{-1})$. Очевидно, що $F(1,\alpha)=0$. Для виконання рівності $F(v,\alpha)$ необхідно, щоб для деякого числа $w\neq 1$ виконувалося

$$F'_v(w,\alpha)=\frac{w^{\alpha}+w^{-\alpha}-w-w^{-1}}{w}=0.$$

Також, $F'_v(w,1)=0$. Отже, має існувати число $\beta\neq 1$ таке, щоб виконувалася рівність

$$F''_{v\alpha}(w,\beta)=\ln w\frac{w^{\beta}-w^{-\beta}}{w}=0.$$

Але, як бачимо, це неможливо, тому функція ρ має єдину точку максимуму $t=q$.

Тепер розглянемо випадок, коли параметр q лежить на межі квадрату $[0,1]^2$. Нехай, наприклад, $q_1=0$, $q_2>0$. Тоді рівності (7), (8) та (10) матимуть вигляд:

$$(1-t_1)^{n_1}(1-t_2)^{n_2}=(1-q_2)^{n_2},$$
$$n_1t_1(1-t_1)^{n_1-1}(1-t_2)^{n_2}+n_2(1-t_1)^{n_1}t_2(1-t_2)^{n_2-1}=n_2q_2(1-q_2)^{n_2-1}.$$

Якщо $t_1 = 0$, то з першого рівняння маємо $t_2 = q_2$. Покажемо, що випадок $t_2 = 0$ неможливий. З попередніх рівностей маємо:

$$(1-t_1)^{n_1} = (1-q_2)^{n_2},$$
$$n_1 t_1 (1-t_1)^{n_1-1} = n_2 q_2 (1-q_2)^{n_2-1}. \tag{13}$$

Або, позначивши $\alpha = n_1/n_2$ та поділивши друге рівняння на перше:

$$(1-t_1)^{\alpha} = 1-q_2,$$
$$\alpha \frac{t_1}{1-t_1} = \frac{q_2}{1-q_2}.$$

Виражаючи q_2 з обох рівнянь, отримаємо

$$1-(1-t_1)^{\alpha} = \alpha t_1 (1-t_1)^{\alpha-1}.$$

Розглянемо функцію $F(t_1, \alpha) = (1-t_1)^{\alpha} + \alpha t_1 (1-t_1)^{\alpha-1}$. Маємо, $F(0,\alpha) = 1$, а її похідна по t_1 дорівнює

$F'_{t_1}(v,\alpha) = -\alpha(1-t_1)^{\alpha-1} + \alpha(1-t_1)^{\alpha-1} - \alpha(\alpha-1)t_1(1-t_1)^{\alpha-2} = -\alpha(\alpha-1)t_1(1-t_1)^{\alpha-2}$.

Вона зберігає свій знак при $t_1 \in (0,1]$, тому рівність $F(t_1,\alpha) = 1$ виконується лише при $t_1 = 0$, а в цьому випадку рівняння (13) не виконується.

Інші випадки розглядаються аналогічно. □

Література

1. А. А. Боровков, *Математическая статистика*, "Наука", Новосибирск, 1997.
2. Р. Є. Майборода, *Оцінка розподілів компонентів сумішей що змінюються*, Укр. мат. журнал. **48** (1996), № 4, 562–566.
3. Р. Є. Майборода, *Статистичний аналіз сумішей*, ВПЦ "Київський університет", Київ, 2003.
4. Р. Є. Майборода, О. В. Сугакова, *Оцінювання та класифікація за спостереженнями із суміші*, ВПЦ "Київський університет", Київ, 2008.
5. А. М. Щербіна, *Оцінювання середнього у моделі суміші зі змінними концентраціями*, Теорія ймовірностей та математична статистика **84** (2011), 142–154.
6. А. М. Щербіна, *Порівняння оцінок середніх значень для сумішей зі змінними концентраціями на модельованих даних*, Вісник Київ. нац. ун-ту. Математика. Механіка, (2011).
7. O. O. Kubaychuk, *Estimation of moments by observations from mixtures with varying concentrations*, Theory of Stochastic Processes **8**(**24**) (2002) №3–4, 226–232.
8. G. J. McLachan and D. Pell, *Finite Mixture Models*, Wiley, NY, 2000.
9. S. Newcomb, *A generalized theory of the combination of observations so as to obtain the best result*, Amer. J. Math. **8** (1886), №4, 343–366.
10. K. Pearson, *Contribution to the mathematical theory of evolution*, Trans. Roy. Sos. A. **185** (1894), 71–110.
11. A. Shcherbina and R. Maiboroda, *Merging data from anonymous and open surveys: two-population problems*, Proceedings of the Baltic–Nordic–Ukrainian Summer School on Survey Statistics, "TViMS", Kyiv, 2009.
12. D. M. Titterington, A. F. M. Smith, and U. E. Makov, *Statistical Analysis of Finite Mixture Distributions*, Wiley, New York, 1985.

Кафедра теорії ймовірностей, математичної статистики та актуарної математики механіко–математичний факультет, Національний університет ім. Тараса Шевченка, проспект Глушкова, 2, Київ 03127, Україна

Адреса електронної пошти: artshcherbina@gmail.com

Надійшла 20/05/2011