УДК 519.25:53.088

СТАТИСТИЧНА ОБРОБКА ЕМПІРИЧНИХ ДАНИХ З ЗАСТОСУВАННЯМ СУМІШЕЙ ІМОВІРНІСНИХ РОЗПОДІЛІВ

Горошко А.В., Ройзман В.П.
Хмельницький національний університет, Україна

Допустимі значення параметрів з певною надійністю визначаються методами математичної статистики на основі отриманих експериментальних даних. Найчастіше дослідники обробляють емпіричні дані, виходячи із параметричних статистичних гіпотез. Перевага застосування типових законів розподілу (нормального, логарифмічно нормального, експоненціального закону, закону Вейбулла, гамма-розподілу тощо) полягає в їх достатній вивченості та можливості отримання спроможних, незміщених і відносно високоефективних оцінок параметрів. Однак вказані вище типові закони розподілу не володіють необхідним різноманіттям форм, тому їх застосування не дає необхідної загальності подання випадкових величин, які зустрічаються при дослідженні систем. Якщо апроксимація на основі типових розподілів не дає бажаної точності статистичних оцінок, у нагоді може стати непараметричний підхід. Методи непараметричної статистики є досить ефективними у багатьох задачах, які достатньо часто виникають на практиці, коли дослідник обробляє відносно малочисельні вибірки, нічого не знаючи про параметри досліджуваної генеральної сукупності. Одним із недоліків непараметричних критеріїв є низька статистична потужність у порівнянні зі стандартними параметричними критеріями.

Серед всіх задач статистичної оцінки параметрів можна виокремити клас задач, в якому оцінці підлягають емпіричні дані, сформовані під дією декількох домінуючих причин, причому виявити ці причини і розділити вибірку на відповідні до них підвибірки не видається можливим. Зокрема, такі задачі часто виникають на виробництві при статистичній оцінці параметрів деякої вибірки деталей, які потрапили на підприємство із різних партій. Густина розподілу (ГР) імовірностей досліджуваних параметрів може бути полімодальною. В роботі [1] наведені приклади полімодальних гістограм розподілу фізико-механічних характеристик деяких технічних об’єктів та причини появи полімодальності.

Очевидно, що через ненормальність ГР застосування параметричного підходу в цьому разі буде не ефективним.

Непараметричний підхід не дасть відповіді про причини полімодальності і не зможе розкрити внутрішню структуру даних з урахуванням можливої полімодальності законів їх розподілу. Ці проблеми тягнуть за собою труднощі із встановленням допусків експериментально досліджуваних параметрів.

Суть запропонованого авторами методу обробки емпіричних даних, що не підкоряються унімодальним законам розподілу, полягає у представлені і обробці емпіричної ГР у вигляді суперпозиції k функцій ГР f_i з вектором параметрів $\boldsymbol{\theta}_i$ (компонент суміші), $i = 1,2,...,k$, $2 \le k < \infty$ у вигляді

$$f(x) = \sum_{i=1}^{k} \rho_i f_i(x, \boldsymbol{\theta}_i), \quad (1)$$

де $x \in \mathbb{R}$, ρ_i - апріорна імовірність (ваговий коефіцієнт) *i*-ї компоненти суміші, $\rho_i \in (0,1)$, $\sum_{i=1}^{k} \rho_i = 1$. В загальному випадку умова приналежності $\forall i,\ f_i(X, \boldsymbol{\theta}_i)$ до одного параметричного сімейства не ставиться.

Нехай в результаті експерименту одержана вибірка значень $x = \{x_1, x_2..., x_n\}$. Для подальшої обробки результатів експерименту, перш за все, необхідно провести декомпозицію (розщеплення) суміші, тобто визначити невідомі параметри $\rho_1, \rho_2, ...\rho_{i-1}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_k$, наприклад, максимізуючи функцію максимальної правдоподібності [2].

$$W(\rho, \theta, x) = \prod_{j=1}^{n} \sum_{i=1}^{k} \rho_i f_i(x_j, \boldsymbol{\theta}_i), \quad (2)$$

прирівнюючи до нуля її частинні похідні по шуканих параметрах. Як правило, замість пошуку максимуму функції $W(\rho, \theta, x)$ простіше шукати максимум її логарифму

$$\ln W(\rho, \theta, x) = \sum_{j=1}^{n} \log\left(\sum_{i=1}^{k} \rho_i f_i(x_j, \theta_i) \right), \quad (3)$$

але навіть така постановка задачі без застосування спеціальних прийомів викликає значні труднощі. Тому для декомпозиції суміші (1) застосовують спеціальні методи: EM-алгоритм і його модифікації - SEM, CEM, MCEM, SAEM тощо; наближені методи, такі як метод фіксованих компонент з використанням МНК та методу найменшим

модулей, а також Баєсовський класифікатор, описані, наприклад, в роботах [2, 3].

Запропонований авторами метод декомпозиції сумішей базується на апроксимації функції ГР імовірностей функцією типу (1) за допомогою МНК або інтерполяції на деякій точковій множині. В той же час відомо, що емпіричні дані вибірки можуть бути представлені лише варіаційним рядом, гістограмою або емпіричною функцією розподілу імовірностей

$$F_n(x) = \frac{1}{n}\sum_{j=1}^{n} 1\left(x_j < x\right). \qquad (4)$$

Запропонований метод декомпозиції сумішей імовірнісних розподілів можна застосовувати як окремо, так і разом із відомими методами, що підвищує точність і вірогідність знайдених оцінок.

Одержання закону розподілу імовірності досліджуваного параметра у вигляді (1) дозволяє перейти до вирішення однієї із важливих практичних задач - призначення допустимого значення цього параметра з певною надійністю. Введемо деякі обмеження, а саме будемо вважати, що компоненти суміші (1) мають нормальний закон розподілу імовірностей.

Як відомо, розсіювання значень досліджуваного параметра залежить від прийнятого способу виготовлення виробу. Межі інтервалів розсіювання визначаються законами розподілу параметра, який розглядається як випадкова величина, що є сумою випадкових величин, кожна з яких викликається одним з нездоланних чинників. Якщо кількість доданків у сумі досить велике, то може виникнути два варіанти при призначенні функції розподілу параметра.

У разі, коли величина кожної зі складових у описаної раніше сумі мала в порівнянні з її величиною, за центральною граничною теоремою [4] розподіл суми близький до нормального. Фізично це умова малості кожного доданка означає, що жоден з факторів, що обумовили появу відповідної випадкової величини, не має переважаючого значення.

Якщо ж серед зазначених факторів з'являються один або кілька домінуючих, то відповідні доданки мають переважне значення в сумі і закон розподілу суми стає полімодальним.

У разі, якщо отримана гістограма описується полімодальним законом розподілу, подальші дії з призначення допустимого значення досліджуваного параметра можуть здійснюватись двома шляхами.

1. Розглядається підвибірка з мінімальним (максимальним) значенням μ_i. Очевидно, що характеристика цієї підгрупи мінімальна

(максимальна). Отже, визначена характеристика для таких виробів може бути прийнята і для всієї партії, оскільки отримані при цьому похибки підуть у запас міцності. У цьому випадку подальша обробка експериментальних даних може відбуватися тільки для зазначеної нормально розподіленої підвибірки значень з параметрами розподілу μ_i, σ_i, як описано вище.

Якщо є можливість розділити вихідну вибірку виробів на підвибірки, об'єднані однією з домінуючих причин появи розкиду значень, то аналогічні операції з обробки експериментальних даних слід проводити для кожної підвибірки.

2. Визначені параметри дозволяють записати інтегральну функцію розподілу з «вагами»

$$F(x) = \sum_{i=1}^{k} \frac{\rho_i}{\sigma_i \sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2} \right) dx, \qquad (6)$$

яку, як і Гаусову випадкову величину, за допомогою комп'ютера можна задати таблицею наступним чином. Для кожного значення величини x, яке змінюється з певним числовим інтервалом, наприклад, 0,1, за таблицею функції розподілу нормованого нормального розподілу

$$\Phi^x(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{x^2}{2} \right) dx$$

можна визначити імовірність

$$\gamma_i = \frac{1}{\sigma_i \sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2} \right) dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(x-\mu_i)/\sigma_i} \exp\left(-\frac{x^2}{2} \right) dx \ , \ i = 1, 2, ..., k$$

і далі значення інтегральної функції з «вагами»

$$F^x(x) = \sum_{i=1}^{k} \rho_i \gamma_i \ . \qquad (7)$$

Це означає, що функція $F^x(x)$ буде задана таблицею. Отримана таблиця дозволяє не тільки за значеннями x визначати величину функції $F^x(x)$, але і навпаки – за заданими значеннями функції визначати величину аргументу.

Необхідно відзначити, що, по-перше, другий шлях більш точний, оскільки він враховує функції розподілу всіх підвибірок, а по-друге, більш універсальний, адже з його допомогою можна вирішувати поставлену задачу у випадку довільного розподілу, якщо попередньо

скласти для нього таблицю залежності довірчої ймовірності і аргументу інтегральної функції розподілу досліджуваної величини.

Також необхідно відзначити, що другий спосіб призначення допусків при полімодальному розподілі параметра поширюється як частинний випадок і на унімодальний закон. Більше того, призначення допуску за допомогою інтегральної функції розподілу в цьому окремому випадку може слугувати навіть доповненням і уточненням способу розв'язання аналогічної задачі при унімодальних законах розподілу параметра, описаного раніше в припущенні, що істинне значення вимірюваної величини збігається з її математичним сподіванням.

Порівняємо ефективність використання параметричного і непараметричного методів та методу представлення вибірки як суміші гауссіан. Для цього змоделюємо суміш $F(x)$ двох нормальних розподілів $F_1(x)$ і $F_2(x)$ типу $N\left(\mu,\ \sigma^2\right)$, де μ_1=74, σ_1=6,6, μ_2=81, σ_2=5,7, з відповідними вагами у суміші ρ_1=66%, ρ_2=34%. Об'єм вибірки n=21:

$$\left\{\begin{matrix}72,33;\ 77.03;\ 72.05;\ 71.15;\ 71.83;\ 72.53;\ 72.57;\ 74.46;\ 73.49;\ 75.51;\\ 71.81;\ 76.06;\ 70.12;\ 76.25;\ 80.42;\ 81.40;\ 76.31;\ 77.97;\ 83.81;\ 85.84;\ 80.34\end{matrix}\right\}. \quad (8)$$

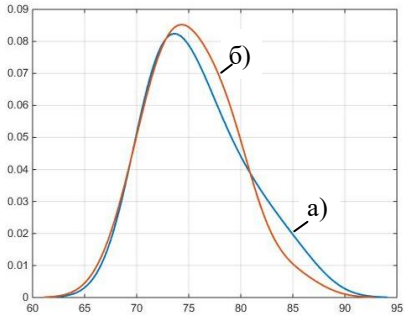

Рис. 1. Графіки функцій густини розподілу:
а) початкової суміші імовірнісних розподілів $F(x)$
б) суміші після декомпозиції її складових $F(x)$

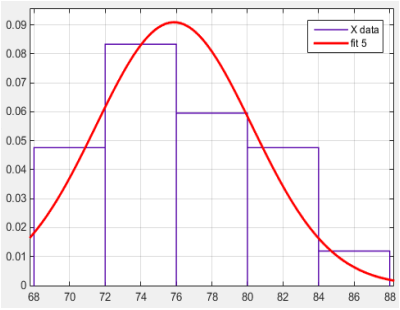

Рис. 2. Графік нормального розподілу, що описує імовірнісний розподіл вибірки (10).

Гістограма розподілу суміші, представлена на рис. 1 (а), «нагадує» гістограму розподілу закону Гаусса. Тести Лілліефорса і Ярки-Бера на непротиріччя розподілу генеральної сукупності значень

випадкової величини нормальному закону показали позитивний результат. Параметри нормального закону, який може описати досліджувану вибірку, дорівнюють μ=75,9, σ=4,4, де μ - точкова оцінка математичного сподівання, σ - точкова оцінка середнього квадратичного відхилення (рис. 2). При цьому значення від'ємного логарифму функції правдоподібності дорівнює -60,35. Для порівняння, значення від'ємного логарифму функції правдоподібності для закону Вейбулла дорівнює -62,90, тобто закон Вейбулла краще описує досліджувану вибірку.

Після декомпозиції суміші типу (6) за допомогою ЕМ-алгоритму, одержані наступні її характеристики

$$F(x)=\sum_{i=1}^{2}\frac{\rho_i}{\sigma_i\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right)dx,$$

μ_1=73,6, σ_1=5,0, ρ_1=71%, μ_2=81,3, σ_2=8,6, ρ_2=29%. (9)

Графік функції густини імовірності суміші $F(x)$, що складається з двох законів Гауса з ідентифікованими параметрами (9), представлений на рис. 1 (б).

Внаслідок похибок декомпозиції (відновлення), одержана густина розподілів дещо відрізняється від початкової. Значення від'ємного логарифму функції правдоподібності суміші дорівнює -88,78. Отже, не зважаючи на правомірність апроксимації вибірки (8) нормальним розподілом, суміш нормальних розподілів в 1,5 рази краще описує досліджувану вибірку даних.

Побудована таблиця значень інтегральної характеристики (7) показала, що значення імовірності попадання випадкової величини у заданий інтервал для суміші (9) точніше, ніж нормального розподілу.

Література

1. Горошко А.В. Методи обробки емпіричних даних, що підпорядковуються багатомодальним законам розподілу / А.В. Горошко, В.П. Ройзман // Вісник Хмельницького національного університету. -2013. №4. -С. 195-201.
2. Прикладная статистика: Классификация и снижение размерности: Справ. изд. / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин; Под ред. С. А. Айвазяна. – М.: Финансы и статистика, 1989. – 607 с.: ил.
3. S.F. Nielsen. The stochastic EM algorithm: estimation and asymptotic results. – *Bernoulli, 2000, vol. 6, No. 3, p. 457-489.*
4. Вентцель Е.С. Теория вероятностей / Е.С. Вентцель - М.: Наука, 1969. - 576с.