УДК 519.24:519.652.3

П.О.Приставка, О.Г.Чолишкіна

ОЦІНКА ПАРАМЕТРА ЕКСПОНЕНЦІАЛЬНОГО РОЗПОДІЛУ У ВИПАДКУ МАЛОЇ ВИБІРКИ

В роботі запропоновано спосіб оцінки параметра експоненціального розподілу при обмеженій кількості даних експерименту. Показано переваги способу у порівняні з методом моментів на основі оцінки середнього арифметичного вибірки.

Постановка проблеми. Питання про можливість аналізу після обробки обмежених обсягів статистичних даних є неоднозначним та суперечливим. Ефективність та спроможність оцінок на основі малих вибірок забезпечити практично неможливо, але, в той же час, потреба в результатах аналізу виникає в багатьох випадках практичної діяльності: дослідження рідкісних явищ, оцінка терміну активного існування високо надійних технічних систем, діагностика за певними видами захворювань, тощо. Тому винесення рекомендацій по технологіям опрацювання малих вибірок, зокрема визначення нових підходів до оцінювання функції розподілу ймовірностей, було та залишається питанням актуальним.

Аналіз публікацій. В теорії перевірки статистичних гіпотез про однорідність вибірок питання винесення висновків на основі малих обсягів даних традиційно вирішується на основі t-тесту, f-тесту, критеріїв дисперсійного аналізу, непараметричних критеріїв (Манна-Уітні, *H*-критерій, тощо). Обґрунтуванню зазначених критеріїв приділяли увагу відомі вчені: У.Госсет (t-розподіл Стьюдента), Р.Фішер, Дж.Снедекор та інш. Для обробки малих обсягів дво- та багатовимірних даних, з метою встановлення наявності зв'язку між окремими ознаками, можна використовувати рангові коефіцієнти Спірмена, Кенделла, коефіцієнти таблиць сполучень.

Задача оцінки функції розподілу ймовірностей на разі малої кількості даних безпосередньо стосується винесення висновків про властивості генеральної сукупності, тож обсяг вибірки в першу чергу є визначальним з точки зору її репрезентативності.

У припущенні нормального закону теоретичної функції розподілу ймовірності випадкової величини, оцінка математичного сподівання при

обсягах даних від 2 до 10 вирішується в середньому адекватно на підставі оцінки середнього арифметичного вибірки. Для оцінки середньоквадратичного σ досить використання формул, що наведено в таблиці [1, стор.18], наприклад, при обсягах 3 та 4 даних мають місце співвідношення:

$$\sigma = 0,5908\left(t_3 - t_1\right),$$

$$\sigma = 0,4539\left(t_4 - t_1\right) + 0,1102\left(t_3 - t_2\right),$$

де t_i, $i = \overline{1,4}$ – елементи варіаційного ряду.

Питанню непараметричного оцінювання функції розподілу ймовірності неперервної випадкової величини присвячено відому монографію Д.В.Гаспарова та В.І.Шаповалова [2]. Показано, що оцінка емпіричної функції розподілу ймовірності, отримана на підставі методу вкладень та його модифікацій, є стійкою відносно статистичних характеристик вихідного ряду.

В авторських роботах [3; 4] оцінку функції щільності розподілу за малими вибірками запропоновано вирішувати на основі локальних поліноміальних сплайнів, близьких до інтерполяційних у середньому, на основі *B*-сплайнів [5; 6]. На основі моделювання з параметричних розподілів вибірок обмеженого обсягу (15 та більше даних) показано [4] переваги зазначених сплайнів при оцінюванні в середньому перших чотирьох моментів у порівнянні з методами прямокутних вкладень та апріорно-емпіричних функцій. Проте, варто відзначити, що не вдається на такому обсязі даних досягти прийнятної якості оцінювання при суттєвій лівій асиметрії теоретичних функцій щільності розподілу.

Поставимо за мету даної роботи провести дослідження питання оцінки експоненціального розподілу при обсягах даних до 10 елементів. Актуальними такі дослідження можуть бути при вирішенні задач теорії надійності та при розробці моделей систем масового обслуговування, зокрема, для оцінювання інтенсивностей потоків вимог та обслуговування.

Виклад основного матеріалу. Нехай за рівномірним розбиттям $\Delta_h : t_i = ih$ $\left(\tilde{\Delta}_h : t_i = \left(i + 0,5\right)h\right)$, $i \in Z$, $h > 0$ вісі реалізацій випадкової величини $\xi(\omega)$, на підставі вибірки $\Omega_{1,N} = \left\{t_l ; l = \overline{1,N}\right\}$ проведено гістограмну оцінку, а отже одержано F_{1,N_i}, $t \in \left[t_i ; t_{i+1}\right)$, $i \in Z$ – масив емпіричної оцінки функції розподілу, тобто, будемо вважати, що на підставі $\Omega_{1,N}$ одержано масив

$$\Delta_h : \left\{t_i, F_{1,N_i} ; i = \overline{0,m-1}\right\},$$

де m – кількість класів при гістограмній оцінці.

Тоді найпростішим, з точки зору реалізації у програмному забезпеченні обробки даних, при оцінюванні функції розподілу $F(t)$ для $\forall t \in [0; t_{\max}]$ (тут $t_{\max}$ – максимальне із зафіксованих спостережень) буде [6] застосування сплайну $S_{3,0}(F_{1,N}, t)$:

$$S_{3,0}(F_{1,N}, t) = \frac{1}{48}\left(-F_{1,N_{i-1}} + 3F_{1,N_i} - 3F_{1,N_{i+1}} + F_{1,N_{i+2}}\right)x^3 +$$
$$+\frac{1}{16}\left(F_{1,N_{i-1}} - F_{1,N_i} - F_{1,N_{i+1}} + F_{1,N_{i+2}}\right)x^2 +$$
$$+\frac{1}{16}\left(-F_{1,N_{i-1}} - 5F_{1,N_i} + 5F_{1,N_{i+1}} + F_{1,N_{i+2}}\right)x +$$
$$+\frac{1}{48}\left(F_{1,N_{i-1}} + 23F_{1,N_i} + 23F_{1,N_{i+1}} + F_{1,N_{i+2}}\right),$$

де $x = \frac{2}{h}(t - t_i) - 1$; $i = \left[\frac{t}{h}\right] + 1$; $[\square]$ – ціла частина.

Зважаючи на властивості функції розподілу, в якості невизначених значень $F_{1,N_{-2}}$, $F_{1,N_{-1}}$, F_{1,N_m}, достатньо узяти такі:

$$F_{1,N_{-2}} = 0, \quad F_{1,N_{-1}} = 0, \quad F_{1,N_m} = 1.$$

Для оцінювання функції експоненціального розподілу ймовірностей

$$F(t;\lambda) = \begin{cases} 0, & -\infty < t < 0, \\ 1 - \exp(-\lambda t), & 0 \le t < \infty, \end{cases} \qquad (1)$$

згідно методу моментів, достатньо визначити оцінку параметр λ так:

$$\hat{\lambda} = \frac{1}{\bar{t}},$$

де

$$\bar{t} = \sum_{i=1}^{N} t_i \Big/ N \qquad (2)$$

– оцінка математичного сподівання за вибіркою $\Omega_{1,N}$. Для сплайн-оцінювання математичного сподівання можна скористатись квадратурною формулою (наприклад лівих прямокутників) для чисельного обрахунку за виразом для теоретичного моменту:

$$\nu_1 = \int_0^{\infty} t\, dF(t),$$

де, в якості оцінки функції $F(t)$ виступає сплайн $S_{3,0}\left(F_{1,N},t\right)$:

$$\hat{\nu}_1=\sum_{j=1}^{n}t_j\left(S_{3,0}\left(F_{1,N},t_j+\Delta t\right)-S_{3,0}\left(F_{1,N},t_j\right)\right), \qquad (3)$$

де t_j, $j=\overline{1,n}$ – деякі значення з інтервалу $\left[0;t_{\max}\right]$, узяті з рівномірним кроком Δt, досить малим, щоб забезпечити прийнятну точність вразу (2).

Іншим чином оцінку математичного сподівання можна отримати із (1). Зокрема, в точці оцінки теоретичного моменту $\tilde{\nu}_1^*\in\left[0,t_{\max}\right)$, згідно оцінювання параметра λ а основі методу моментів, має виконуватись:

$$F\left(\hat{\nu}_1^*,\frac{1}{\hat{\nu}_1^*}\right)=1-\exp\left(-\frac{1}{\hat{\nu}_1^*}\cdot\hat{\nu}_1^*\right),$$

або

$$F\left(\hat{\nu}_1^*,\frac{1}{\hat{\nu}_1^*}\right)=1-\exp(-1),$$

звідки оцінка $\hat{\nu}_1^*$ визначається як квантиль оцінки функції розподілу при рівні ймовірності $\alpha=1-\exp(-1)$, тобто:

$$\hat{\nu}_1^*=\hat{F}^{-1}\left(\hat{\nu}_1^*,\ 1\Big/\hat{\nu}_1^*\right), \qquad (4)$$

де в якості оцінки $\hat{F}\left(\hat{\nu}_1^*,\ 1\Big/\hat{\nu}_1^*\right)$ пропонується використати непараметричну оцінку на основі сплайну $S_{3,0}\left(F_{1,N},t\right)$. Визначення величини $\hat{\nu}_1^*$ на основі сплайн-оцінки неважко зробити, наприклад на основі методу ділення відрізку навпіл.

Проведемо експериментальні дослідження оцінок (2)-(4) з метою визначення їх властивостей при малих обсягах даних вибірки $\Omega_{1,N}$. Опис схеми експерименту такий.

Крок 1. Випадковим чином визначаємо кількість даних вибірки від 4-х до 9-ти: $N\in\left[4;9\right]$.

Крок 2. Моделюємо дані вибірки $\Omega_{1,N}$, розподілені за експоненціальним законом розподілу ймовірностей згідно виразу:

$$t_l = \frac{1}{\lambda}\ln\frac{1}{1-random}, \quad l = \overline{1,N},$$

де λ – параметр розподілу (для визначеності $\lambda = 5$);

$random$ – рівномірно розподілене псевдовипадкове число з інтервалу $[0;1)$.

Крок 3. Для змодельованої вибірки $\Omega_{1,N}$ проводиться формування варіаційного ряду та розбиття його на класи, причому кількість класів m гістограмної оцінки визначимо так:

$$m = \begin{cases} 3, & N \in [4;6], \\ 4 & N \in [7;9]. \end{cases}$$

Крок 4. Обраховуємо за зберігаємо для обробки оцінки (2)-(4).

Крок 5. Кроки 1 – 4 повторюємо 2000 разів для отримання представницького обсягу вбіркових статистик для подальшого аналізу.

Результати проведеного експерименту зведено до таблиці (табл.1) та наведено на графіках (рис.1 – 4).

Таблиця 1.

Значення оцінок математичного сподівання
за результатами експерименту

Оцінка	Середнє оцінки	Медіана оцінки	Теоретичне значення
$\overline{t}$	0,1241	0,1198	0,2
$\hat{\nu}_1$	0,12689	0,12246	
$\hat{\nu}_1^*$	0,17092	0,16304	

Зауважимо, що потреба в залучені до аналізу медіани оцінки математичного сподівання за результатами експерименту викликана асиметричністю функції щільності вибіркового розподілу статистик (2)-(4) що продемонстровано на графіках (рис.1 – 3).

Як видно з таблиці (табл.1), прийнятно оцінити математичне сподівання на основі виразів (2) та (3) в середньому неможливо. Поясненням цього є те, що для моделі експоненціального розподілу поява реалізації випадкової величини $\xi(\omega)$ на «хвості» розподілу – малоймовірна подія і для малих обсягів, що є предметом дослідження, зустрічається не часто, тобто, маємо заниження оцінки у порівнянні з теоретичним моментом. Проте, навіть, коли така подія має місце, це автоматично впливає на оцінку, шляхом її суттєвого завищення (правий хвіст функції щільності на графіках).

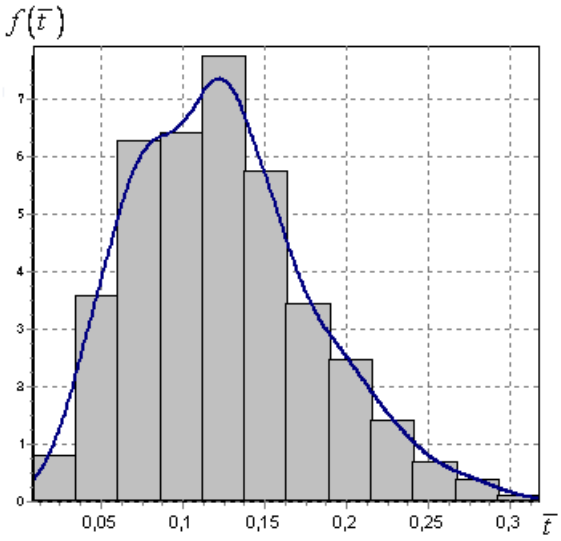

Рис.1. Функція щільності та нормована гістограма оцінки (2) за результатами експерименту

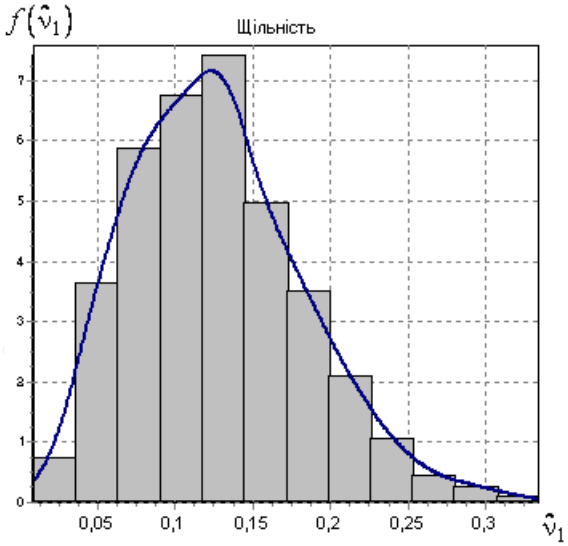

Рис.2. Функція щільності та нормована гістограма оцінки (3) за результатами експерименту

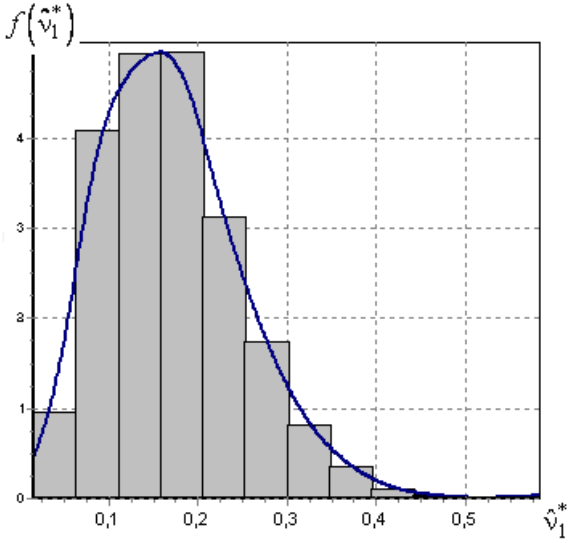

Рис.3. Функція щільності та нормована гістограма оцінки (4) за результатами експерименту

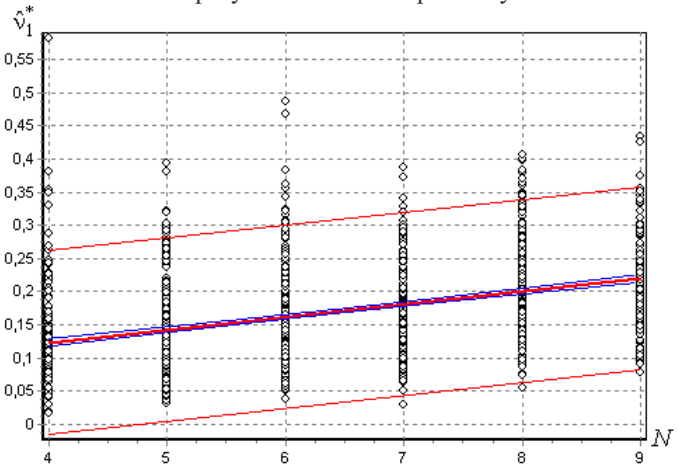

Рис.4. Регресія з довірчим інтервалом та толерантними межами оцінки $\hat{\nu}_1^*$, залежно від обсягу даних вибірки

Для оцінки за виразом (4) ситуація порівняно краща. Приблизно чверть значень оцінок, отриманих в ході експерименту, розташовано в межах від 0,17 до 0,23 і майже половина значень в межах від 0,14 до 0,26.

Якщо проаналізувати залежність оцінювання за виразом (4) від обсягу даних вибірки (рис.4), то видно, що регресія оцінки $\hat{\nu}_1^*$, залежно від обсягу даних вибірки має лінійну модель і для кількості від 7-ми до 9-ти практично відповідає теоретичному значенню математичного сподівання 0,2.

Висновки. На основі результатів експериментів з імітаційного моделювання (табл.1) показано, що при оцінюванні параметра λ в моделі експоненціального розподілу за методом моментів спостерігається завищення такої оцінки. Пов'язана така ситуація з тим, що в малих вибірках значення з «хвостів» експоненціального розподілу зустрічаються нечасто, отже, оцінка математичного сподівання, що є оберненою величиною оцінки $\hat{\lambda}$, у більшості випадків експериментів є заниженою (рис.1).

Таким чином, якщо в якості оцінки математичного сподівання обирати оцінку середнього вибірки, результат буде не адекватним. Аналогічна ситуація при використанні оцінки теоретичного моменту на основі непараметричної сплайн-оцінки функції розподілу (рис.2). Більш адекватним може бути оцінка на основі виразу (4), яка побудована на оцінці кванти лі, що отримана за використанням локального поліноміального сплайну а основі *B*-сплайнів третього порядку, близького до інтерполяційного у середньому (рис.3 – 4).

Подальша дослідження можуть полягати в дослідженні запропонованого підходу при обробці даних в задачах моделювання систем масового обслуговування та при вирішенні задач оцінки терміну активного існування високо надійних технічних систем.

Література.

1. Бабак В.П., Білецький А.Я., Приставка О.П., Приставка П.О. Статистична обробка даних. – К.: "МІВВЦ", 2001. – 388 с.
2. Гаскаров Д.В., Шаповалов В.И. Малая выборка. – М.: Статистика, 1978. –248 с.
3. Приставка Ф.А. Непараметрическая оценка плотности вероятностей по малым выборкам / Придніпровський науковий вісник. – 1998. – N101 (168). – С. 25 – 29.
4. Приставка П.О., Смойловська О.О. Обробка вибірок обмеженого обсягу з використанням поліноміальних сплайнів / Актуальні проблеми автоматизації та інформаційних технологій : Зб. наук. праць. – Д.: Навчальна книга, 2001.–Т.4. – С. 86 – 95.
5. Лигун А.А., Шумейко А.А. Асимптотические методы восстановления кривых. – К.: ІМ НАН України, 1996. – 358 с.
6. Приставка П.О. Поліномаільні сплайни при обробці даних – Д.: Вид-во Дніпропетр. ун-ту, 2004. – 236 с.