УДК 519.234.2:004.67

О.М. Мацуга, Г.О. Басова

Дніпропетровський національний університет ім. Олеся Гончара

ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ ВІДНОВЛЕННЯ РОЗПОДІЛІВ ЗА МАЛИМИ ВИБІРКАМИ

Розроблено програмне забезпечення «SmallSample» відновлення розподілів за малими вибірками. Проведено порівняння роботи непараметричних методів відновлення розподілів за малими вибірками.

Ключові слова: *мала вибірка, програмне забезпечення, відновлення, функція розподілу, функція щільності.*

Разработано программное обеспечение «SmallSample» восстановления распределений по малым выборкам. Проведено сравнение работы непараметрических методов восстановления распределений по малым выборкам.

Ключевые слова: *малая выборка, программное обеспечение, восстановление, функция распределения, функция плотности.*

The program «SmallSample» for distributions restoration using small samples is developed. The distributions nonparametric restoration methods for small samples are compared.

Key words: *small sample, program, restoration, distribution function, density function.*

Постановка проблеми. В умовах обмеженості інформації часто доводиться мати справу з малими вибірками, під якими розуміються вибірки з малою кількістю спостережень над випадковою величиною. Більшість статистичних методів (як параметричних, так і непараметричних) малоефективні у таких випадках, оскільки вони асимптотичні. Тому існує необхідність у використанні методів, спеціально розроблених для вибірок малого обсягу. Такі методи запропоновано [1], проте їх програмна реалізація недоступна. У зв'язку з цим актуальна задача розробки програмного забезпечення, орієнтованого на статистичну обробку малих вибірок. У роботі увага приділена одному з етапів статистичної обробки – відновленню розподілів за малими вибірками.

Аналіз останніх досліджень і публікацій. Історично одними з перших робіт в області статистичного аналізу малої вибірки можна

вважати роботи Стюдента і Фішера, в яких досліджувались задачі оцінювання характеристик випадкової величини. Потім у 60–70-х роках XX століття над цією проблемою працювало багато інших математиків, таких як А.Н. Колмогоров, А.А. Петров, Л.Н. Большев, І.Н. Володін та інші [1].

Пік основних публікації стосовно задачі оцінювання функції розподілу ймовірностей випадкової величини за вибіркою малого обсягу припадає на 70–80-ті роки XX ст. Уперше дана задача була розглянута у роботі В.В. Чавчанідзе, В.А. Кумсішвілі. Результати цього та подальших досліджень узагальнено наприкінці 80-х років XX ст. у монографії Д.В. Гаскарова та В.І. Шаповалова [1], яку можна вважати основною роботою з питань статистичної обробки малих вибірок.

Серед запропонованих методів оцінювання функції розподілу випадкової величини за малою вибіркою можна виділити методи прямокутних вкладів, зменшення невизначеності, апріорно-емпіричної функції, які можуть бути застосовані навіть у випадку вибірки з 3–5 елементів [1]. Для них характерний індивідуальний підхід до кожної окремої реалізації вибірки, що відрізняє їх від методів побудови оцінок за великими вибірками. В якості додаткової апріорної інформації передбачається знання інтервалу $[a;b]$, на якому змінюється випадкова величина.

Оцінювання функції розподілу по зазначеним методам проводиться наступним чином.

Нехай спостереження над дійсною неперервною випадковою величиною X задано у вигляді вибірки $\left\{x_i; i=\overline{1,N}\right\}$. За вибіркою побудовано варіаційний ряд $\left\{x_i, k_i, p_i; i=\overline{1,n}\right\}$, де $x_1 < x_2 < \ldots < x_n$; n – кількість варіант; x_i – значення i-ї варіанти; k_i – частота i-ї варіанти; $\sum_{i=1}^{n} k_i = N$; $p_i = \frac{k_i}{N}$ – відносна частота i-ї варіанти; $\sum_{i=1}^{n} p_i = 1$.

Метод прямокутних вкладів (МПВ) дозволяє знайти оцінку функції щільності розподілу ймовірностей випадкової величини за вибіркою у вигляді [1]

$$f^*(x)=\frac{1}{N+1}\left(f_0(x)+\sum_{i=1}^{N}\psi_{x_i}(x)\right),$$

де $f_0(x)$ – апріорна компонента, відносно якої припускається що вона являє собою щільність рівномірного на $[a;b]$ розподілу

$$f_0(x)=\begin{cases}\dfrac{1}{b-a}, & x\in[a;b],\\ 0, & x\notin[a;b];\end{cases}$$

$\psi_{x_i}(x)$ – функція вкладу реалізації x_i довжиною d, яка забезпечує «розмивання» інформації про випадкову величину, яку було отримано від реалізації x_i

$$\psi_{x_i}(x)=\begin{cases}\dfrac{1}{d}, & x\in\left[x_i-\dfrac{d}{2};x_i+\dfrac{d}{2}\right],\\ 0, & x\notin\left[x_i-\dfrac{d}{2};x_i+\dfrac{d}{2}\right].\end{cases}$$

Оптимальне значення величини d залежить від обсягу вибірки та типу оцінюваного розподілу [1].

Оцінка функції розподілу $F^*(x)$ для МПВ одержується шляхом інтегрування функції щільності $f^*(x)$:

$$F^*(x)=\frac{1}{N+1}\left(F_0(x)+\sum_{i=1}^{N}\varphi_{x_i}(x)\right),$$

де

$$F_0(x)=\begin{cases}0, & x<a,\\ \dfrac{x-a}{b-a}, & x\in[a;b],\\ 1, & x>b;\end{cases}$$

$$\varphi_{x_i}(x)=\begin{cases}0, & x<x_i-\dfrac{d}{2},\\ \dfrac{1}{d}\left(x-x_i+\dfrac{d}{2}\right), & x\in\left[x_i-\dfrac{d}{2};x_i+\dfrac{d}{2}\right],\\ 1, & x>x_i-\dfrac{d}{2}.\end{cases}$$

У методі зменшення невизначеності (МЗН) оцінка функції розподілу знаходиться за варіаційним рядом за формулою [1]

$$F^*(x) = \frac{x - x_{i-1}}{x_i - x_{i-1}}\left(F^*(x_i) - F^*(x_{i-1})\right) + F^*(x_{i-1}),$$

коли $x \in [x_{i-1}; x_i]$ та, якщо $x = x_i$, то

$$F^*(x_i) = \frac{1}{N+1}\left(\frac{x_i - a}{b-a} + (i - 0{,}5) + (k_i - 1)\right).$$

За методом апріорно-емпіричної функції (МАЕФ) оцінка функції розподілу визначається виразом [1]

$$F(x) = \omega F_a(x) + (1-\omega) F_N(x),$$

де $F_a(x)$ – апріорно задана функція розподілу; $F_N(x)$ – емпірична функція розподілу; ω – коефіцієнт достовірності інформації про апріорний розподіл.

Оцінка функції щільності $f^*(x)$ у МЗН та МАЕФ може бути одержана шляхом диференціювання $F^*(x)$ за змінною x.

Слід зазначити, що емпірична функція розподілу $F_N(x)$ є найбільш проста та традиційна непараметрична оцінка функції розподілу $F(x)$. Можливість її використання в якості оцінки функції розподілу було доведено В.І. Глівенко та Ф.П. Кантеллі. Емпірична функція розподілу визначається варіаційним рядом за формулою

$$F_N(x) = \begin{cases} 0, & x \le x_1, \\ \sum\limits_{j=1}^{i} p_j, & x \in (x_i; x_{i+1}], \\ x, & x > x_n. \end{cases}$$

Розглянуті вище методи є непараметричні.

Спеціальних параметричних методів для оцінювання функції розподілу за малими вибірками не запропоновано. Застосування ж традиційних параметричних методів (максимальної правдоподібності, найменших квадратів, моментів тощо [2]) потребує знання мінімального обсягу вибірки, за якого вони можуть бути застосовані. Існуючі дослідження з цього приводу малочисельні та розрізнені. Деякі автори вважають обмеженим вибірки обсягом менше 200 елементів, інші називають малими вибірки в менш ніж 50 елементів [1]. Тому поставлена у роботі задача передбачає програмну

реалізацією не лише непараметричних методів оцінювання функцій розподілу за малими вибірками, а й параметричних з оцінкою мінімального обсягу вибірки.

Постановка задачі. Задано вибірку спостережень $\left\{x_i; i=\overline{1,N}\right\}$ над дійсною неперервною величиною X з невідомою функцією розподілу $F(x)$. Припускається, що вибірка малого обсягу. Ставиться задача за вибіркою знайти оцінку $F^*(x)$ функції розподілу $F(x)$, або, що тотожно, знайти оцінку $f^*(x)$ щільності розподілу $f(x)$.

Для вирішення цієї задачі необхідно розробити програмне забезпечення, яке б дозволяло здійснювати непараметричне оцінювання функцій розподілу і щільності розподілу за МПВ, МЗН, МАЕФ, побудувати емпіричну функцію розподілу, а також проводити параметричне відновлення розподілів за класичними методами та оцінювати мінімальний обсяг вибірки, за якого параметричні методи забезпечують задану точність відновлення.

Основний матеріал. Для вирішення поставленої задачі створено програмне забезпечення «SmallSample» у середовищі Microsoft Visual Studio 2010 на мові програмування C#.

Структурними елементами програмного забезпечення є:

1. Обчислювальне ядро.
2. Блок роботи з даними, який забезпечує завантаження, моделювання та збереження вибірок.
3. Блок візуалізації, який реалізує представлення результатів у вигляді таблиць і графіків, а також забезпечує зручний інтерфейс, надаючи можливість аналізувати результати.

Ядро програмного забезпечення дозволяє проводити:

1. Непараметричне оцінювання функції розподілу класичним методом, тобто шляхом побудови емпіричної функції розподілу.
2. Непараметричне оцінювання функцій розподілу та щільності розподілу ймовірностей випадкової величини на основі МПВ, МЗН, МАЕФ [1].
3. Параметричне оцінювання функцій рівномірного, нормального, експоненціального, Релея та Вейбулла розподілів на основі методів максимальної правдоподібності (ММП), найменших квадратів (МНК), моментів (ММ) [2]. З метою ідентифікації передбачена побудова ймовірнісного паперу для кожного з вказаних типів розподілів.
4. Визначення мінімального обсягу вибірки, за якого параметричні методи забезпечують задану точність оцінювання функції розподілу.

5. Моделювання вибірок з розподілів рівномірного, нормального, експоненціального, Релея та Вейбулла з метою тестування зазначених вище методів оцінювання на вибірках малого обсягу [3].

6. Обчислення максимальної різниці між оцінкою функції розподілу та теоретичною (змодельованою) функцією.

Загальна схема роботи програми представлена на рис. 1.

Рис. 1. Блок схема роботи програми «SmallSample»

Створений програмний продукт має головне вікно, на якому розміщений блокнот із сімома закладками:

1. «Моделювання даних». На цій закладці можна задати параметри розподілу, який моделюється, обсяг вибірки та переглянути у таблиці змодельовану вибірку (рис. 2).

2. «Вар. ряд; ряд, розбитий на класи». Тут міститься таблиця з варіаційним рядом та рядом розбитим на класи, а також графік гістограми. На закладці можна задати кількість класів.

3. «Непараметричне відновлення». Тут міститься ще один блокнот, кожна закладка якого відповідає певному методу непараметричного

відновлення розподілу (класичний, МПВ, МЗН, МЕАФ). Для кожного методу на графіки виводяться оцінки функцій розподілу та щільності розподілу, а також теоретичні функції у разі моделювання (рис. 3)

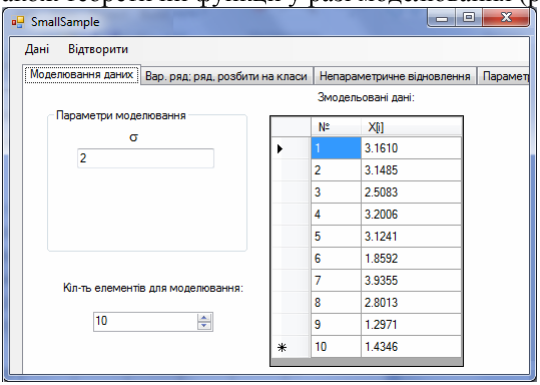

Рис. 2. Вигляд закладки «Моделювання даних»

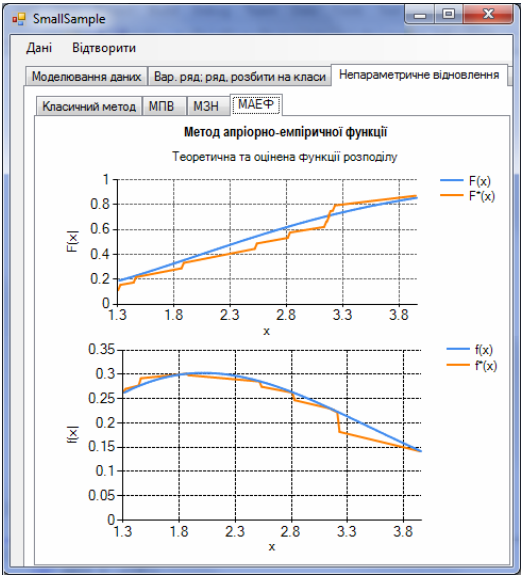

Рис. 3. Вигляд закладки «Непараметричне відновлення»

4. «Параметричне відновлення». На закладці розміщено таблицю, в яку виводяться знайдені оцінки параметрів та у разі моделювання даних значення параметра моделювання і його відхилення від оцінки (рис. 4). Також на закладці знаходиться графік з ймовірнісним

папером (будується папір для того типу розподілу, який обрано для відновлення) та графіки з функціями розподілу і щільності розподілу (оціненими та, у разі моделювання даних, теоретичними).

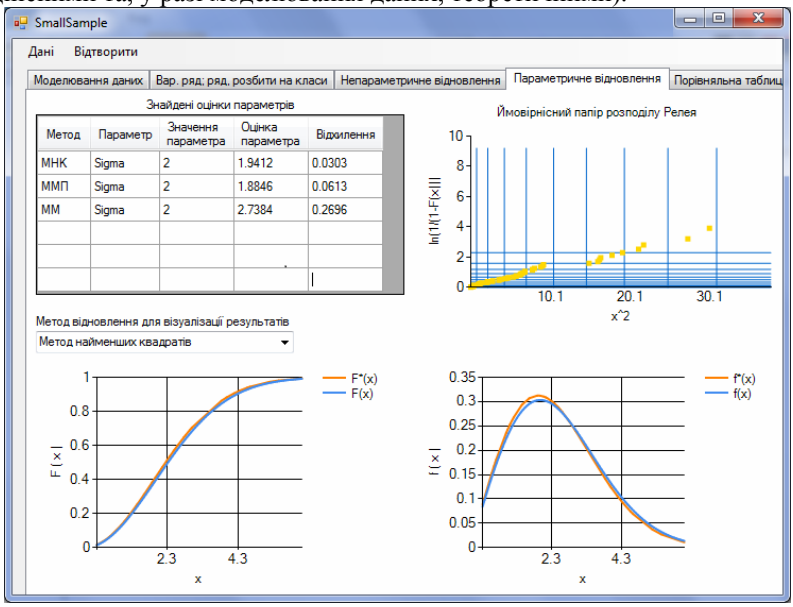

Рис. 4. Вигляд закладки «Параметричне відновлення»

5. «Порівняльна таблиця». Тут розміщена таблиця, до якої виводиться максимальна різниця між оціненими та теоретичними функціями розподілу і щільності розподілу. Таблиця заповнюється лише у випадку, коли вибірка була змодельована.

6. «Експеримент 1». На закладці при натисненні відповідної кнопки до таблиці виводяться результати наступного експерименту. Моделюється задана кількість вибірок вказаного обсягу з певного розподілу. За кожною вибіркою усіма непараметричними методами оцінюється функція розподілу, знаходиться максимальна різниця між оціненою функцією розподілу та теоретичною. За масивом максимальних різниць обчислюються середнє арифметичне та середньоквадратичне, які і виводяться у таблицю.

В якості прикладу нижче наведено результати одного з експериментів, під час яких моделювалось по 100 вибірок обсягом 5 з кожного розподілу (табл. 1). Через брак місця результати

представлено частково, лише для рівномірного, нормального та експоненціального розподілів.

Таблиця 1

Результати експерименту 1

Тип розподілу	Класичний метод	МПВ	МЗН	МАЕФ
Середнє арифметичне максимальної різниці				
Рівномірний	0,304	0,23	0,2	0,17
Нормальний	0,301	0,17	0,20	0,18
Експоненціальний	0,298	0,22	0,19	0,19
Середньоквадратичне максимальної різниці				
Рівномірний	0,120	0,095	0,077	0,060
Нормальний	0,109	0,087	0,073	0,063
Експоненціальний	0,118	0,110	0,072	0,058

Дані таблиці 1 свідчать, що найкращі результати дозволяє одержати метод апріорно-емпіричної функції. Висновок підтверджується і на інших типах розподілів та інших обсягах вибірок.

7. «Експеримент 2». На цій закладці при натисненні кнопки до таблиці виводяться результати такого експерименту. Для заданого типу розподілу моделюється задана кількість вибірок різного обсягу (мінімальний, максимальний обсяг та крок, з яким він має змінюватися, задаються). За кожною вибіркою здійснюється відновлення розподілу трьома параметричними методами і обчислюється максимальна різниця між оціненою функцією розподілу та теоретичною. За масивом максимальних різниць розраховується середнє, яке і виводяться до таблиці.

В якості прикладу наведено результати експериментів, під час яких моделювались по 100 вибірок з експоненціального та Релея розподілів (табл. 2). У таблиці через брак м'ясця представлено результати лише для п'яти обсягів. Одержані дані свідчать, що зі збільшенням обсягу вибірки максимальна різниця між оціненою та теоретичною функціями розподілу зменшується, що цілком природно. Задавшись точністю в 0,01, можна визначити мінімальний обсяг вибірки, за якого середнє значення максимальної різниці не перевищує цю точність. Для експоненціального розподілу ця умова починає виконуватися для вибірки обсягом 30 елементів. Для Релея у разі застосування методу максимальної правдоподібності мінімальний обсяг вибірки також 30 елементів, але для інших методів він вищий –

100 елементів. Результати експериментів на вибірках з різних розподілів засвідчили, що під час застосування параметричних методів оцінювання розподілу, щоб максимальна різниця оціненої та теоретичної функцій розподілу була менше 0,01, обсяг вибірки має перевищувати 70–100 елементів.

Таблиця 2

Результати експерименту 2

Тип розподілу	Обсяг вибірки	ММП	МНК	ММ
Експоненціальний	10	0,0304	0,0518	0,0304
	15	0,0590	0,0445	0,0597
	25	0,0589	0,0267	0,0534
	30	0,0095	0,0076	0,0087
	50	0,0058	0,0048	0,0058
Релея	10	0,0340	0,0522	0,3616
	15	0,0323	0,0343	0,3305
	25	0,0167	0,0135	0,2567
	30	0,0093	0,0154	0,1120
	50	0,0052	0,0168	0,1871

Висновки. Створено програмне забезпечення «SmallSample» для відновлення розподілів за малими вибірками. Програмне забезпечення пройшло ретельне тестування на даних імітаційного моделювання, результати якого дозволяють рекомендувати його для використання у практичних задачах. Результати тестування методів непараметричного відновлення розподілів засвідчили, що найбільш точно оцінити функцію розподілу дозволяє метод апріорно-емпіричної функції.

Бібліографічні посилання

1. **Гаскаров Д.В.** Малая выборка / Д.В. Гаскаров, В.И. Шаповалов. – М., 1978. – 248 с.

2. **Бабак В.П.** Статистична обробка даних / В.П. Бабак, А.Я. Білецький, О.П. Приставка, П.О. Приставка. – К., 2001. – 388 с.

3. **Байбуз О.Г.** Системи масового обслуговування / О.Г. Байбуз, О.П. Приставка, П.О. Приставка. – Д., 2001. – 84 с.

Надійшла до редколегії 21.06.2012