УДК 519.233.2:519.254

О. М. Мацуга, А. С. Черкашина

Дніпропетровський національний університет імені Олеся Гончара

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ РОБАСТНОГО ВІДНОВЛЕННЯ РОЗПОДІЛІВ

Розроблено інформаційну технологію робастного відновлення розподілів, яку реалізовано у вигляді програмного забезпечення «RobustProcedre». Результати тестування програми на даних імітаційного моделювання підтвердили адекватність технології.

Ключові слова: *інформаційна технологія, робастне оцінювання, відновлення розподілу, сплайн-розподіл.*

Разработана информационная технология робастного восстановления распределений, которую реализовано в виде программного обеспечения «RobustProcedre». Результаты тестирования программы на данных имитационного моделирования подтвердили адекватность технологии.

Ключевые слова: *информационная технология, робастное оценивание, восстановление распределения, сплайн-распределение.*

The information technology of distribution robust restoration is developed. It was realized as software «RobustProcedre». Software testing results on modeling data corroborate the technology.

Key words: *information technology, robust estimate, distribution restoration, spline-distribution.*

Постановка проблеми. Під час обробки і аналізу вибіркових даних значної уваги потребують аномальні спостереження, значення яких різко відрізняються від інших. Такі спостереження можуть суттєво впливати на результати статистичного дослідження, зокрема результати відновлення розподілу ймовірностей. Для того щоб зменшити їх вплив, доцільно застосовувати робастні методи математичної статистики. Такі методи добре вивчені, проте їх програмна реалізація у автоматизованих системах обробки даних поки що недостатня.

Аналіз останніх досліджень і публікацій. Вперше теоретичний підхід до проблеми робастності в статистиці був запропонований Хьюбером [1], який ввів так звані «*M*-оцінки», що є узагальненням оцінок максимальної правдоподібності. Суттєво інший підхід, заснований на функціях впливу, був запропонований Хампелем [2]. Стосовно задачі відновлення розподілів *M*-оцінки більш гнучкі, оскільки допускають пряме узагальнення на багатопараметричний випадок.

Постановка задачі. Нехай результати спостережень задано у вигляді вибірки $\left\{x_i; i=\overline{1,N}\right\}$, де N – кількість спостережень; x_i – спо-

стережуване значення в *i*-му експерименті. Припускається, що вибірка може містити аномальні значення. За вибірковими даними потрібно відновити розподіл ймовірностей $F\left(x;\vec{\Theta}\right)$, де $\vec{\Theta}=\left\{\theta_1,\theta_2,\ldots,\theta_p\right\}$ – вектор параметрів розподілу, і знайти оцінку $\hat{\vec{\Theta}}=\left\{\hat{\theta}_1,\hat{\theta}_2,\ldots,\hat{\theta}_p\right\}$, стійку відносно можливих відхилень статистичного розподілу від теоретичного.

Ставиться задача розробити інформаційну технологію робастного відновлення розподілів, яку реалізувати у вигляді сучасного програмного забезпечення. В якості моделей розподілів розглянути «чисті» та сплайн-розподіли з класу нормального, логарифмічно нормального, експоненціального та Вейбулла [3, 4].

Основний матеріал. Обчислювальна технологія робастного відновлення розподілів містить такі етапи:

1. Проведення первинного статистичного аналізу вибіркових даних, який передбачає побудову варіаційного ряду, розбиття ряду на класи, побудову гістограми та емпіричної функції розподілу, обчислення статистичних характеристик вибірки [3].

2. Ідентифікація розподілу з класу нормального, логарифмічно нормального, експоненціального та Вейбулла. Ідентифікацію пропонується проводити за допомогою процедури автоматизованої ідентифікації [5].

3. Робастне оцінювання параметрів ідентифікованого розподілу за нижченаведеною процедурою.

4. Перевірка вірогідності відновлення на основі будь-якого відомого критерію згоди, наприклад Колмогорова або Мізеса [3].

Реалізація нижченаведеної процедури робастного оцінювання параметрів розподілу потребує зведення функції розподілу $F\left(x;\vec{\Theta}\right)$ до лінійного вигляду $z\left(t;\vec{\Theta}\right)$, де $t=\phi\left(x\right)$ – перетворення над показником x, $z\left(t\right)=\varphi\left(F\left(x\right)\right)$ – перетворення над функцією розподілу і переформування масиву $\left\{x_i,F_N\left(x_i\right);i=\overline{1,n}\right\}$, де $F_N\left(x\right)$ – значення емпіричної функції розподілу, *n* – кількість варіант, у масив $\left\{t_i,z_i;i=\overline{1,n-1}\right\}$. Для розподілів, що розглядаються у роботі, справедливі такі перетворення:

– нормального:

$$\varphi\left(F\left(x\right)\right)=\Phi^{-1}\left(F\left(x\right)\right),\qquad \phi\left(x\right)=x,$$

де $\Phi^{-1}(\square)$ – обернена до функції Лапласа;

– логарифмічно нормального:

$$\varphi\left(F\left(x\right)\right)=\Phi^{-1}\left(F\left(x\right)\right),\qquad \phi\left(x\right)=\ln x;$$

– експоненціального:

$$\varphi\left(F\left(x\right)\right)=\ln\frac{1}{1-F\left(x\right)},\qquad \phi\left(x\right)=x;$$

– Вейбулла:

$$\varphi\left(F\left(x\right)\right) = \ln\ln\frac{1}{1-F\left(x\right)}, \qquad \phi\left(x\right) = \ln x.$$

«Чисті» розподіли, які описуються лише одним параметром (як експоненціальний), зводяться до лінійного вигляду

$$z\left(t\right) = \theta_1 t\,.$$

Лінійний вигляд сплайн-розподілів з k вузлами склеювання з класу однопараметричних розподілів можна подати виразом

$$z\left(t\right) = \begin{cases} \theta_1 t, & t \le t^{(1)}, \\ \left(\theta_1 - \theta_2\right)t^{(1)} + \theta_2 t, & t^{(1)} \le t \le t^{(2)}, \\ \dots \\ \sum\limits_{j=1}^{m}\left(\theta_j - \theta_{j+1}\right)t^{(j)} + \theta_{m+1}t, & t^{(m)} \le t \le t^{(m+1)},\ m = 1,\dots,k-1, \\ \dots \\ \sum\limits_{j=1}^{k}\left(\theta_j - \theta_{j+1}\right)t^{(j)} + \theta_{k+1}t, & t \ge t^{(k)}, \end{cases}$$

де $t^{(m)}$ – m-й вузол склеювання.

«Чисті» двопараметричні розподіли зводяться до лінійного виду

$$z\left(t\right) = \theta_1 + \theta_2 t\,.$$

Сплайн-розподіли з k вузлами склеювання з класу двопараметричних розподілів зводяться до лінійного вигляду

$$z\left(t\right) = \begin{cases} \theta_1 + \theta_2 t, & t \le t^{(1)}, \\ \theta_1 + \left(\theta_2 - \theta_3\right)t^{(1)} + \theta_3 t, & t^{(1)} \le t \le t^{(2)}, \\ \dots \\ \theta_1 + \sum\limits_{j=1}^{m}\left(\theta_{j+1} - \theta_{j+2}\right)t^{(j)} + \theta_{m+2}t, & t^{(m)} \le t \le t^{(m+1)},\ m = 1,\dots,k-1, \\ \dots \\ \theta_1 + \sum\limits_{j=1}^{k}\left(\theta_{j+1} - \theta_{j+2}\right)t^{(j)} + \theta_{k+2}t, & t \ge t^{(k)}. \end{cases}$$

Процедуру знаходження робастної оцінки вектора параметрів розподілу можна подати у наступному вигляді:

1. Покладається $q = 0$ – номер ітерації і знаходиться початкове наближення оцінки вектора параметрів розподілу $\hat{\hat{\Theta}}^{(0)}$ з умови мінімуму функціоналу залишкової дисперсії

$$S^2 = \sum_{i=1}^{n-1} w_i^{(q)} \left(z_i - z\left(t_i; \hat{\hat{\Theta}}^{(q)} \right) \right)^2, \qquad (1)$$

де $w_i^{(0)} = 1$, $i = \overline{1, n-1}$; $z\left(t_i; \hat{\hat{\Theta}} \right)$ – значення лінеаризованої функції розподілу; z_i – емпіричне значення лінеаризованої функції розподілу.

У разі відновлення «чистого» розподілу мінімізація функціоналу (1) реалізує ідею методу найменших квадратів [3]. Знаходження мінімуму функціоналу (1) для сплайн-розподілу передбачає додатково оцінювання вузлів склеювання шляхом перебору [3]. Для цього вузли склеювання послідовно поміщаються в один з варіантів, і визначаються оцінки інших параметрів сплайн-розподілу шляхом мінімізації функціоналу (1). Серед усіх варіантів вузлів та відповідних їм оцінок обираються ті, для яких значення функціоналу (1) максимальне.

2. Збільшується номер ітерації $q = q + 1$.

3. Обчислюються залишки

$$\varepsilon_i = z_i - z\left(t_i; \hat{\hat{\Theta}}^{(q-1)} \right), \; i = \overline{1, n-1}.$$

4. Визначається масштабний множник

$$c = \frac{1}{0,6745} \operatorname{median} \left| \varepsilon_i - \operatorname{median}\left(\varepsilon_i \right) \right|.$$

5. Знаходяться значення $y_i = \varepsilon_i / c$, $i = \overline{1, n-1}$.

6. Використовуючи поточне наближення оцінки $\hat{\hat{\Theta}}^{(q-1)}$ та обраний критерій, знаходяться ваги

$$w_i^{(q)} = \frac{\psi\left(y_i \right)}{y_i}, \; i = \overline{1, n-1},$$

де $\psi\left(y \right) = \dfrac{\partial \rho\left(y \right)}{\partial y}$; $\rho\left(y \right)$ – функція критерію.

Під час реалізації технології використано такі запропоновані у літературі критерії [6]:

– метод найменших квадратів:

$$\rho\left(y \right) = \frac{1}{2} y^2, \; w\left(y \right) = 1, \; y \in \left(-\infty; +\infty \right);$$

– Хьюбера:

$$\rho(y)=\begin{cases}\frac{1}{2}y^2, y\in[-a;a]\\ a|y|-\frac{1}{2}a^2, y\notin[-a;a]\end{cases},\ w(y)=\begin{cases}1, y\in[-a;a]\\ \frac{a}{|y|}, y\notin[-a;a]\end{cases},$$

де $a > 0$ – точка зламу;

– Рамсея:

$$\rho(y)=\frac{1}{a^2}\left(1-e^{-a|y|}\left(1+a|y|\right)\right),\ w(y)=e^{-a|y|},\ y\in(-\infty;+\infty),$$

де $a > 0$ – точка зламу;

– хвиля Ендрюса:

$$\rho(y)=\begin{cases}a\left(1-\cos\left(\frac{y}{a}\right)\right), y\in[-a\pi;a\pi]\\ 2a, y\notin[-a\pi;a\pi]\end{cases},\ w(y)=\begin{cases}\sin\left(\frac{y}{a}\right), y\in[-a\pi;a\pi]\\ 0, y\notin[-a\pi;a\pi]\end{cases},$$

де $a\pi > 0$ – точка зламу;

– Тьюкі:

$$\rho(y)=\begin{cases}\frac{1}{2}y^2-\frac{y^4}{4a^2}, y\in[-a;a]\\ \frac{1}{4}a^2, y\notin[-a;a]\end{cases},\ w(y)=\begin{cases}1-\frac{y^2}{a^2}, y\in[-a;a]\\ 0, y\notin[-a;a]\end{cases},$$

де $a > 0$ – точка зламу;

– Гампеля:

$$\rho(y)=\begin{cases}\frac{1}{2}y^2, y\in[-a;a]\\ a|y|-\frac{1}{2}a^2, y\in[-b;-a]\ \text{або}\ y\in[a;b]\\ a\frac{c|y|-\frac{1}{2}y^2}{c-b}-\frac{7a^2}{6}, y\in[-c;-b]\ \text{або}\ y\in[b;c]\\ a(b+c-a), y\in[-\infty;-c]\ \text{або}\ y\in[c;+\infty]\end{cases},$$

$$w(y)=\begin{cases}1, y\in[-a;a]\\ \dfrac{a}{|y|}, y\in[-b;-a] \text{ або } y\in[a;b]\\ a\dfrac{c/|y|-1}{c-b}, y\in[-c;-b] \text{ або } y\in[b;c]\\ 0, y\in[-\infty;-c] \text{ або } y\in[c;+\infty]\end{cases},$$

де $c > b > a > 0$ – точки зламів.

7. З умови мінімуму функціоналу (1), використовуючи ваги, знайдені на кроці 6, обчислюється наступне наближення оцінки вектора параметрів $\hat{\bar{\Theta}}^{(q)}$.

8. Перевіряється умова

$$\max_{j=1,p}\left|\hat{\theta}_j^{(q)}-\hat{\theta}_j^{(q-1)}\right|<e.$$

Якщо вона виконується, то $\hat{\bar{\Theta}}^{(q)}=\left\{\hat{\theta}_1^{(q)},\hat{\theta}_2^{(q)},\ldots,\hat{\theta}_p^{(q)}\right\}$ приймаються за шукану оцінку вектора параметрів, інакше здійснюється перехід на крок 2.

Обчислювальна технологія робастного оцінювання параметрів розподілу склала ядро програмного забезпечення «RobustProcedre», яке реалізовано у середовищі Borland Delphi 7.0.

Тестування створеної інформаційної технології проведено на даних імітаційного моделювання. Експерименти полягали у моделюванні вибірки за певним законом розподілу ймовірностей, додаванні до вибірки аномальних значень із подальшим відновленням розподілу за допомогою інформаційної технології. Нижче, у якості прикладу, наведено результати одного з експериментів.

Під час експерименту моделювались дані зі сплайн-розподілу Вейбулла з одним вузлом склеювання. Параметри моделювання та знайдені оцінки параметрів подано у таблиці 1. На ймовірнісному папері розподілу Вейбулла нанесено оцінки лінеаризованої функції розподілу, знайдені методом найменших квадратів та за допомогою описаної вище технології (рис. 1). Функція, що проходить нижче на графіку, оцінена за допомогою створеної технології.

Імовірність вірогідного відновлення розподілу методом найменших квадратів за критерієм Колмогорова складає 0,89, а робастного відновлення – 0,95.

Таблиця 1

Параметри та знайдені оцінки параметрів сплайн-розподілу Вейбулла з одним вузлом

Параметр	Значення під час моделювання	Оцінка методом найменших квадратів	Робастна оцінка
α	3	2,93	2,92
β_1	0,8	0,79	0,80
β_2	5	5,57	5,08
x_0	2,5	2,53	2,51

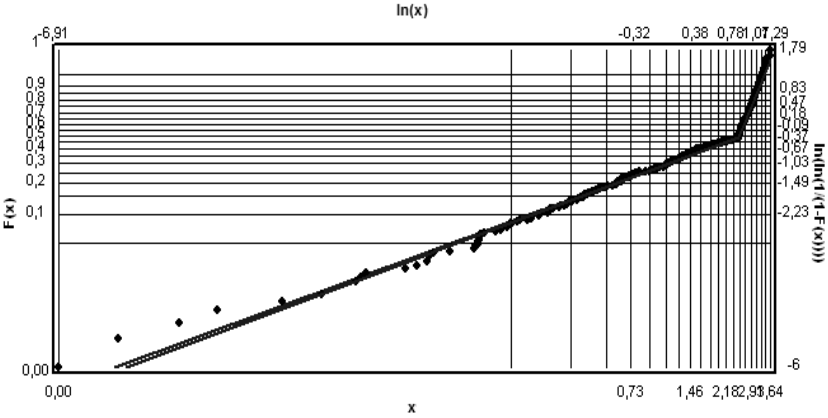

Рис. 1. Ймовірнісний папір розподілу Вейбулла з відновленою лінеаризованою функцією розподілу

Висновки за результатами проведеної роботи:

− Було створено інформаційну технологію робастного відновлення розподілів з класів нормального, логарифмічно нормального, експоненціального та Вейбулла («чистих» розподілів та сплайн-розподілів з одним і двома вузлами).

− Інформаційну технологію реалізовано у вигляді програмного забезпечення «RobustProcedre».

− Проведено тестування технології на модельованих даних.

− Результати тестування свідчать про наступне: 1) оцінки параметрів, знайдені на основі запропонованої технології, більш близькі до істинних оцінок параметрів (параметрів моделювання), ніж оцінки, знайдені методом найменших квадратів; 2) робастне відновлення

більш вірогідне, ніж відновлення на основі методу найменших квадратів, за критерієм згоди Колмогорова.

Бібліографічні посилання

1. **Хьюбер П.** Робастность в статистике / П. Хьюбер. – М. : Мир, 1984. – 304 с.

2. Робастность в статистике. Подход на основе функций влияния / Ф. Хампель, Э. Рончетти, П. Рауссеу, В. Штаэль. – М. : Мир, 1989. – 512 с.

3. Статистична обробка даних / В. П. Бабак, А. Я. Білецький, О. П. Приставка, П. О. Приставка. – К. : МІВВЦ, 2001. – 388 с.

4. **Приставка О. П.** Сплайн-розподіли у статистичному аналізі / О. П. Приставка. – Д. : Вид-во ДДУ, 1995. – 152 с.

5. **Мацуга О. М.** Обчислювальні схеми ідентифікації сплайн-розподілів за ймовірнісним папером / О. М. Мацуга, Г. С. Шубіна // Актуальні проблеми автоматизації та інформаційних технологій. – 2012. – Т. 16. – С. 112–123.

6. **Дрейпер Н.** Прикладной регрессионный анализ / Н. Дрейпер, Г. Смит. – 3-е изд. – М. : Вильямс, 2007. – 912 с.

Надійшла до редколегії 24.06.14