Ivanna BUBELA¹

OPRACOWANIE WYNIKÓW LOSOWYCH OBSERWACJI Z PŁASKO-NORMALNYM ROZKŁADEM METODĄ STATYSTYK POZYCYJNYCH

W artykule zastosowano metodę statystyk pozycyjnych do opracowania obserwacji z płasko-normalnym rozkładem prawdopodobieństwa, który jest splotem rozkładów normalnego i jednostajnego. Zaproponowana metoda zapewnia mniejszą niepewność standardową wyniku w porównaniu do niepewności wartości średniej. Skuteczność metody zwiększa się, gdy wzrasta zawartość składnika jednostajnego. Realizacja metody nie wymaga skomplikowanych obliczeń. Przedstawiono wyniki badań metody - niepewność standardową w zależności od wzajemnej zawartości składników oraz liczby obserwacji.

Słowa kluczowe: niepewność, metoda statystyk pozycyjnych, macierz kowariacyjna, metoda Monte-Carlo.

1. Wstęp

Niezbędnym elementem pomiarów jest ocena niepewności uzyskanych wyników [1]. Często losowe obserwacje pomiarowe nie podlegają żadnemu „czystemu” rozkładowi (na przykład normalnemu, Laplace’a, jednostajnemu, etc.), dlatego należy zastosować rozkład, który jest splotem dwóch lub więcej typowych rozkładów. Przykładowo splotem dwóch typów rozkładu jest rozkład płasko-normalny. Kształt rozkładu sumy $x = x_1 + x_2$ dwóch niezależnych zmiennych losowych x_1 i x_2 zależy od wartości tych zmiennych oraz od ich rozkładów $p_1(x_1)$ i $p_2(x_2)$, a funkcja gęstości rozkładu opisana jest zależnością:

$$p(x) = \int_{-\infty}^{\infty} p_1(x - x_2) \cdot p_2(x_2) dx_2 = \int_{-\infty}^{\infty} p_1(x_1) \cdot p_2(x - x_1) dx_1. \quad (1)$$

¹ Ivanna Bubela, doktorantka Narodowego Uniwersytetu „Politechnika Lwowska”, Ukraina, Lviv, ul. Stepan Bandera 12, 79013, 097-612-58-29, popovych.i@ukr.net

W szczególności, jeśli jedna ze zmiennych losowych, na przykład x_1 ma rozkład normalny z wartością oczekiwaną m_n i odchyleniem standardowym σ_n , a druga zmienna losowa x_2 ma rozkład jednostajny z wartością oczekiwaną m_r i zakresem zmienności a , wtedy rozkład $x = x_1 + x_2$ może być opisany wzorem [2, 3, 4]:

$$p_{n,r}(x, m, a, \sigma) = \frac{F_n((x - m_x + a)/\sigma_n) - F_n((x - m_x - a)/\sigma_n)}{2a}, \quad (2)$$

gdzie $F_n(x)$ - funkcja Laplace'a; $m_x = m_n + m_r$ - wartość oczekiwana sumy. Rozkład prawdopodobieństwa (2) wykorzystywany jest przy obliczaniu niepewności wyników pomiarów metodami przybliżonymi [2-5] i nazywany jest „rozkładem płasko-normalnym” [2, 3].

Podstawą do wyznaczania najlepszych estymatorów z obserwacji o rozkładzie (2) jest znajomość zawartości składowych normalnej i jednostajnej w ich sumie. W praktyce zwykle nie ma takiej informacji a priori, wtedy do wyznaczania najlepszego estymatora z minimalną niepewnością standardową wyniku może być zastosowana metoda statystyk pozycyjnych [6-8].

2. Metoda statystyk pozycyjnych

Zgodnie z tą metodą zasada wyznaczania najlepszych estymatorów położenia $\hat{\mu}$ i szerokości $\hat{\sigma}$ badanej próby losowej polega na wstępnym uporządkowaniu obserwacji: $X_s = \text{sort}(X) = x_{(1)}, x_{(2)}, \dots, x_{(n)}$, a następnie na minimalizacji sumy kwadratów S_R^2 odchyłeń $v_k = x_{ref_k}' - x_{(k)} = \hat{\mu} + x_{ref_k} \cdot \hat{\sigma} - x_{(k)}$, ($m=0$, $\sigma=0$) od obserwacji referencyjnych x_{ref_k} , odpowiadających wybranym rozkładom prawdopodobieństwa [6-8].

Parametry $\hat{\mu}$ i $\hat{\sigma}$ badanej próby obliczane są metodą ważonych najmniejszych kwadratów, która w zapisie macierzowym ma postać [7-9]:

$$(\hat{\mu}, \hat{\sigma})^T = (A^T \cdot W \cdot A)^{-1} A^T \cdot W \cdot X_s = REC \cdot X_s, \quad (3)$$

gdzie: $A^T = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_{ref_1} & x_{ref_2} & \dots & x_{ref_n} \end{pmatrix}$ - macierz obserwacji referencyjnych,

$W = \text{COV}^{-1}$ - macierz wagowa, będąca odwrotnością macierzy kowariancji COV statystyk pozycyjnych, $REC = (A^T \cdot W \cdot A)^{-1} A^T \cdot W$ - macierz rekonstrukcyjna.

Elementy macierzy kowariacyjnej COV są obliczane według wzoru [7-10]:

$$Cov_{k,l} = \iint_{x_j > x_k} s \cdot z \cdot p_{2_{k,l}}(s, z) ds dz - xref_k xref_l, \quad (4)$$

gdzie

$$p_{2_{k,l}}(s, z) = C(n, k, l) \cdot [F(s)]^{k-1} [F(z) - F(s)]^{l-k-1} [1 - F(z)]^{n-l} p(s) p(z) \quad (5)$$

jest łącznym rozkładem prawdopodobieństwa statystyk pozycyjnych k -tej (s) i l -tej (z) [10]:

$$C(n, k, l) = \frac{n!}{(n-l)!(l-k-1)!(k-1)!}. \quad (6)$$

Jeśli typ rozkładu a priori nie jest znany, to w celu wyznaczania najlepszych estymatorów $\hat{\mu}$ i $\hat{\sigma}$ uporządkowane obserwacje wejściowe porównywane są z zestawem J ($j=1, 2, \dots, J$) obserwacji referencyjnych $xref_{1,j}, xref_{2,j}, xref_{3,j}, \dots, xref_{n,j}$, odpowiadających przewidywanej gęstości rozkładów $p_1(x), p_2(x), p_j(x), \dots, p_J(x)$ zmiennej losowej opisanej rozkładem (2) przy różnych proporcjach parametrów a i σ_n . Następnie poszukuje się najlepszej pary nieznanych parametrów $\hat{\mu}, \hat{\sigma}$ z uwzględnieniem warunku najlepszego dopasowania zarejestrowanych obserwacji do obserwacji referencyjnych, to znaczy z warunku minimalnej wartości sumy kwadratów odchyleń resztkowych:

$$\min \left[S_{R,j}^2 = \frac{\left(X_s - A_j \cdot (\hat{\mu}_j, \hat{\sigma}_j)^T \right)^T \cdot W_j \cdot \left(X_s - A_j \cdot (\hat{\mu}_j, \hat{\sigma}_j)^T \right)}{n-2} \right]. \quad (7)$$

Podstawowym problemem praktycznej realizacji przedstawionej wyżej metody jest złożoność obliczania macierzy kowariancji COV (zależności od (3) do (5)); oprócz tego dokładność obliczeń zmniejsza się wraz ze wzrostem liczby obserwacji n . Ten problem jest ważny, gdy liczba obserwacji i wartości składnika jednostajnego oraz normalnego mogą się zmieniać.

Celem prowadzonych badań jest:

- analiza możliwości zastosowania uproszczonej metody statystyk pozycyjnych;
- analiza skuteczności zaproponowanej metody i opracowanie badań statystyczną metodą Monte-Carlo.

3. Znormalizowany rozkład sumy zmiennych losowych o rozkładzie normalnym i jednostajnym

Aby zastosować metodę statystyk pozycyjnych podczas wyznaczania obserwacji referencyjnych, należy najpierw utworzyć macierz kowariancji, która używana jest wyraz z rozkładem unormowanej gęstości zmiennej losowej $y = (x - m_x)/\sigma_x$ z $m_x = 0$, $\sigma_x = 1$. W tym przypadku odchylenie standardowe sumy dwóch składników powinno być równe 1: $\sigma_x = \sqrt{\sigma_n^2 + \sigma_r^2} = \sqrt{\sigma_n^2 + a^2/3} = 1$. Jeśli stosunek odchyleń standardowych składowej normalnej i jednostajnej oznaczmy jako $b = \sigma_n/\sigma_r = \sigma_n\sqrt{3}/a$, wtedy z warunku $\sigma_x = 1$ i zależności (2) otrzymujemy wartości parametrów rozrzutu obydwu rozkładów:

$$\text{dla normalnego } a = \frac{\sqrt{3}}{\sqrt{1+b^2}}, \text{ oraz dla jednostajnego } \sigma_n = \frac{b}{\sqrt{1+b^2}}. \quad (8)$$

Wówczas znormalizowany rozkład sumy składników (2) zależy tylko od jednego parametru b i opisywany jest zależnością:

$$p_{n,r}(y,b) = \frac{\sqrt{1+b^2}}{2\sqrt{3}} \left[F\left(y, \frac{\sqrt{3}}{\sqrt{1+b^2}}, \frac{b}{\sqrt{1+b^2}}\right) - F\left(y, -\frac{\sqrt{3}}{\sqrt{1+b^2}}, \frac{b}{\sqrt{1+b^2}}\right) \right]. \quad (9)$$

Na rysunku 1 przedstawiono rozkład (9) dla różnych wartości parametru b .

Rys. 1. Gęstość rozkładu płasko-normalnego dla różnych wartości stosunku odchyleń standardowych (parametru b): $b=20$; 1 ; $1/2$; $1/20$

Fig. 1. The density Flatten-Gaussian distribution for different ratio standard deviation of (parameter b): $b=20$; 1 ; $1/2$; $1/20$

4. Określanie macierzy próbek referencyjnych

4.1. Obserwacje referencyjne

Do realizacji metody statystyk pozycyjnych potrzebne są obserwacje referencyjne ze znormalizowanym rozkładem (9) i dystrybuantą

$$F_{n,r}(y) = \int_{-\infty}^y p_{n,r}(y, b) dy .$$

Na rysunku 2 pokazane są obserwacje referencyjne

dla kilku wartości liczby obserwacji $n = 21, 31, 41$ i 51 i różnych wartości parametru b tj: $b=20; 1; 1/2; 1/20$. Jak wynika z rysunku obserwacje referencyjne przy $b \geq 1/20$ praktycznie są zbieżne z obserwacjami mającymi rozkład normalny, a przy $b \geq 20$ praktycznie odpowiadają obserwacjom posiadającym rozkład jednostajny.

Rys. 2. Przykłady zestawów referencyjnych próbek ($n = 21, 31, 41$ i 51) odpowiadających płasko-normalnemu rozkładowi przy różnej wartości $b = 20; 1; 1/2; 1/20$

Fig. 2. Examples of sets of sample observations ($n=21, 31, 41$ and 51) corresponding to the density Flatten-Gaussian distribution with a different ratio of $b=20; 1; 1/2; 1/20$

4.2. Macierz kowariancyjna

W celu uproszczenia obliczeń macierzy kowariancji zaproponowano wykorzystanie asymptotycznego przybliżenia dla wariancji i współczynników korelacji pomiędzy dwoma statystykami pozycyjnymi [7-10]. Do wyprowadzenia tych wzorów wykorzystano właściwości kwantyli $x_{(\lambda_1)}$ i $x_{(\lambda_2)}$ rzędu λ_1 i λ_2 próby

prostej, pobranej z populacji o rozkładzie $p(x)$ (dystrybuanta $F(x)$), które przy dużych n mają rozkład asymptotycznie normalny o parametrach [9-10]:

$$m_1 = x_{(\lambda,1)}, \quad m_2 = x_{(\lambda,2)} \\ \sigma_1^2 \approx \frac{\lambda_1(1-\lambda_1)}{n(p(x_{(\lambda_1)}))^2}, \quad \sigma_2^2 \approx \frac{\lambda_2(1-\lambda_2)}{n(p(x_{(\lambda_2)}))^2}, \quad \rho_{1,2} \approx \sqrt{\frac{\lambda_1(1-\lambda_2)}{\lambda_2(1-\lambda_1)}}, \quad (10)$$

Dla $1 \leq k \leq n$ wartości λ_k są równe $\lambda_k = k/(n+1)$ i w tym przypadku $x_{ref_{j;k}} \approx x_{(\lambda,j;k)} = qF_j(\lambda_k) = F_j^{-1}(\lambda_k)$. Wtedy na podstawie zależności (10) przybliżone wartości współczynników macierzy kowariancji mogą być obliczone z zależności:

$$Cov_{k-1,l-1} \approx \rho_k \cdot \sigma_k \cdot \sigma_l = \frac{1}{n(n+1)^2} \begin{cases} \frac{k \cdot (n-l+1)}{p_{n,r}(x_{(k-1)}, b) \cdot p_{n,r}(x_{(l-1)}, b)}, & l \geq k \\ \frac{l \cdot (n-k+1)}{p_{n,r}(x_{(k-1)}, b) \cdot p_{n,r}(x_{(l-1)}, b)}, & l < k \end{cases}. \quad (11)$$

5. Wyniki badań

Aby przetestować skuteczność proponowanej metody przeprowadzono badania z wykorzystaniem metody Monte-Carlo. Przyjęto następujące wartości parametrów: liczba symulacji $M = 10^5$; liczba obserwacji n wynosiła kolejno: 21, 31, 41 i 51 przy stosunku odchyłeń standardowych b równym: 20; 1; 1/2; 1/20; wartość oczekiwana $m_x = 5$, odchylenie standardowe $\sigma_x = 0,1$.

Na podstawie obliczonej macierzy (11) dla każdego rozkładu określano macierze REC_j (3) i wyznaczano wartości parametrów położenia $\hat{\mu}_j$ i szerokości $\hat{\sigma}_j$ oraz ich standardowe niepewności ze wzorów:

$$u_A(\hat{\mu}_j) = \sqrt{d_{0,0_j}^2 \cdot S_{R_j}^2}, \quad u_A(\hat{\sigma}_j) = \sqrt{d_{1,1_j}^2 \cdot S_{R_j}^2}, \quad (12)$$

gdzie $d_{0,0_j}^2$, $d_{1,1_j}^2$ są elementami diagonalnymi macierzy D_j :

$$D_j = (A_j^T \cdot W_j \cdot A_j)^{-1} = \begin{pmatrix} d_{0,0_j}^2 & 0 \\ 0 & d_{1,1_j}^2 \end{pmatrix}, \quad (13)$$

$b=20$	0,02068	0,02182	0,02319	0,01507	0,01543	0,01736
$b=1$	0,01977		0,02283	0,01420		0,01597
$b=1/2$	0,01745		0,02123	0,01205		0,01307
$b=1/20$	0,01335		0,01506	0,00872		0,00734
Liczba obserwacji $n=31$						
$b=20$	0,01768	0,01796	0,01858	0,01266	0,01270	0,01390
$b=1$	0,01702		0,01842	0,01199		0,01276
$b=1/2$	0,01532		0,01718	0,01032		0,01037
$b=1/20$	0,01052		0,01151	0,00672		0,00566
Liczba obserwacji $n=41$						
$b=20$	0,01558	0,01562	0,01598	0,01106	0,01104	0,01187
$b=1$	0,01510		0,01583	0,01053		0,01089
$b=1/2$	0,01382		0,01470	0,00916		0,00889
$b=1/20$	0,00901		0,00950	0,00566		0,00492
Liczba obserwacji $n=51$						
$b=20$	0,01405	0,01400	0,01423	0,00992	0,00990	0,01055
$b=1$	0,01367		0,01411	0,00947		0,00965
$b=1/2$	0,01263		0,01302	0,00828		0,00787
$b=1/20$	0,00804		0,00823	0,00500		0,00450
Notatka: 1 - stosunek odchyłeń standardowych składowej normalnej do prostokątnej; 2, 5 - średnia niepewność standardową; 3, 6 - teoretyczna wartość niepewności; 4, 7 - odchylenie standardowe błędu.

Rys. 3. Zależności wartości średnich niepewności standardowej i teoretycznych wartości niepewności parametrów położenia (a) i szerokości (b) od liczby obserwacji $n=21, 31, 41$ i 51 dla $b=20; 1; 1/2; 1/20$

Fig. 3. Depending on the mean value of the standard uncertainty and theoretical uncertainties location and width parameters at $b=20; 1; 1/2; 1/20$ of the number of observations $n=21, 31, 41$ and 51

Rys. 4. Zależności odchylenia standardowego błędów i teoretyczne wartości niepewności parametrów położenia (a) i szerokości (b) od liczby obserwacji $n=21, 31, 41$ i 51 dla $b=20; 1; 1/2; 1/20$

Fig. 4. Depending the standard deviation of the error and theoretical uncertainties location and width parameters at $b=20; 1; 1/2; 1/20$ of the number of observations $n=21, 31, 41$ and 51

6. Podsumowanie

W zagadnieniach opracowania wyników losowych obserwacji wartości poszukiwanych parametrów mogą być wyrażone dokładnie (z minimalną standardową niepewnością) tylko w przypadku znajomości rozkładu prawdopodobieństwa tych obserwacji. Metoda statystyk pozycyjnych nie wymaga a priori wiedzy o zawartości składowych mających płasko-normalny rozkład prawdopodobieństwa. Ponieważ metoda ta bazuje na porównaniu danych wejściowych z zestawem wartości referencyjnych, którymi są wartości oczekiwane statystyk pozycyjnych, zapewnia automatyczny wybór najlepszych parametrów położenia i szerokości. Skuteczność metody statystyk pozycyjnych zwiększa się przy wzroście w obserwacjach wejściowych składnika o rozkładzie jednostajnym.

Zaproponowana metoda obliczania macierzy kowariancji, która jest niezbędna do realizacji metody, jest prosta i wystarczająco dokładna. Jest to bardzo ważne gdy liczba obserwacji nie jest znana a priori i nie jest możliwe wcześniejsze przygotowanie macierzy potrzebnych do opracowania danych. Dlatego macierze te należy obliczać w procesie opracowywania obserwacji.

Ocena zaproponowanej metody przeprowadzona przy pomocy metody Monte-Carlo dała pozytywny wynik i potwierdziła jej skuteczność.

Literatura

- [1] Guide of the Expression of Uncertainty in Measurement. International Organisation for Standardisation. Switzerland, 1993, 1995. 2007, s. 1-13.
- [2] Blázquez J, García-Berrocal A, Montalvo C, Balbás M. The coverage factor in a Flatten–Gaussian distribution. *Metrologia*, vol. 45 (2008), s. 503-506.
- [3] Dietrich C. F.: Uncertainty, Calibration and Probability. The Statistics of Scientific and Industrial Measurement. Second Edition 1991. The Adam Hilger Series on Measurement Science and Technology, s. 535.
- [4] Fotowicz P. Metody obliczania współczynnika rozszerzenia w oparciu o spłot rozkładu prostokątnego z normalnym. *PAK* nr 4 (2004), s. 13-16.
- [5] Fotowicz P. Wykorzystanie rozkładu płasko-normalnego przy obliczaniu niepewności pomiaru. *PAK*, nr 6 (2011), s. 595-598.
- [6] Dorozhovets. M. Doslidzhennya zastosuvannya zrazkovykh vybirok dlya otsinyuvannya rezultatu vymiryuvannya ta jeho standartnoyi nepevnosti. *Vidbir i obrobka inform.* Vyp. 28 (104), 2008.
- [7] Dorozhovets. M. Opracowanie wyników obserwacji bazujące na przybliżonej metodzie statystyk pozycyjnych. *Pomiary Automatyka, Kontrola*, nr 6, 2014, s. 391-394.
- [8] Dorozhovets M, Popovych I. Opratsyuvannya rezultativ sposterezhennya na osnovi nablyzhenoho metodu poryadkovykh statystyk. *Vymiryuval'na tekhnika ta metrolohiya.* № 75, 2014, s. 8-12.
- [9] Kendal M.G., Stuart A. The Advanced Theory of Statistics. Charles Griffin & Co. Ltd. London, 1966.
- [10] Fisz M. Probability Theory and Mathematical Statistics. John Willey & Sons, London, 1963.

PROCESSING OF THE OBSERVATIONS RESULTS WITH THE FLATTEN-GAUSSIAN DISTRIBUTION BY THE ORDER STATISTICS METHOD

Summary

In the article the method based on order statistics for processing of the observations, which is the sum of observations with normal and uniform distributions, is proposed. This method gives a smaller standard uncertainty of the result compared to a standard uncertainty of a mean value. The efficiency of the method increases with the content of the uniform component. The implementation of the method does not require complicated calculations. The research results of method, which are the standard uncertainty depending on the relative content of components and the number of observations, are presented.

Keywords: uncertainty, method of order statistics, covariance matrix, Monte-Carlo method

DOI: 10.7862/re.2015.31

Tekst złożono w redakcji: maj 2015

Przyjęto do druku: wrzesień 2015