УДК 621.317

ОПРАЦЮВАННЯ РЕЗУЛЬТАТІВ СПОСТЕРЕЖЕННЯ НА ОСНОВІ НАБЛИЖЕНОГО МЕТОДУ ПОРЯДКОВИХ СТАТИСТИК

© Михайло Дорожовець, Іванна Попович, 2014

Національний університет "Львівська політехніка", кафедра інформаційно-вимірювальних технологій, вул. С. Бандери, 12, 79013, Львів, Україна

Запропоновано наближений метод порядкових статистик для опрацювання випадкових спостережень за апріорі невідомого розподілу ймовірності генеральної сукупності. Використання наближеного методу не потребує складних розрахунків інтегралів і коваріаційна матриця визначається за допомогою простих арифметичних операції. Подано результати наближеного методу і продемонстровано його ефективність.

Предложен приближенный метод порядковых статистик для обработки случайных наблюдений при априори неизвестном распределения вероятности генеральной совокупности. Использование приближенного метода не требует сложных расчетов интегралов и ковариационная матрица определяется с помощью простых арифметических операций. Представлены результаты приближенного метода и продемонстрировано его эффективность.

In the article the approximate method of order statistics for the processing of the random observations of unknown a priori probability density distribution is proposed. Using approximate method does not require complex calculations of integrals, and the covariance matrix is determined using simple arithmetic operations. Presents the results of a study of approximate methods and demonstrated its effectiveness.

1. Вступ. З метою підвищення точності вимірювань широко використовують статистичне опрацювання результатів спостережень [1]. Найкраща оцінка результату і його стандартної непевності типу А великою мірою залежить від моделі густини розподілу генеральної сукупності, з якої отримано вибірку результатів спостережень (вибірка) [1]. Стандартна методика опрацювання серії некорельованих спостережень $x_1, x_2, ..., x_n$ (n – кількість зареєстрованих спостережень) забезпечує мінімальне значення стандартної непевності типу А в результаті оцінки (середнє значення) тільки у випадку нормального розподілу ймовірності генеральної сукупності $p(x)$ або відповідного близького йому розподілу. Загалом густина розподілу генеральної сукупності дає інформацію про частість появи тих чи інших результатів спостережень, а також про їх взаємне розташування. Тобто різним моделям густини розподілу генеральної сукупності відповідає різне взаємне розміщення результатів спостережень на числовій осі.

Відомо, що найточнішими вимірюваннями є вимірювання із безпосереднім порівнянням вимірюваної величини із мірою. Цей принцип можна застосувати також і до статистичного опрацювання результатів спостережень. Однак постає проблема створення «міри» для випадкових спостережень.

2. Методика порядкових статистик. Якщо густина розподілу спостережень істотно відрізняється від нормальної, існують інші параметри положення спостережень, крім середнього значення, в яких стандартна непевність менша від стандартної непевності середнього значення. У [2, 3] наведено метод, оснований на порядкових статистиках, в якому безпосередньо використовується інформація про розподіл $p(x)$. В [4] доведено, що в цьому методі для дисперсії $\mathrm{var}(\hat{m})$ оцінки положення $\hat{m}$ і дисперсії середнього значення $\mathrm{var}(\bar{x})$ існує така нерівність:

$$\mathrm{var}(\hat{m}) \le \frac{s^2}{n} = \mathrm{var}(\bar{x}), \tag{1}$$

де рівність $\hat{m} = \bar{x}$ справджується тільки для нормального розподілу.

Однак необхідною умовою для визначення найкращого результату з найменшою стандартною непевністю є точно відомий розподіл ймовірності спостереження $p(x)$. Якщо розподіл не відомий, але відомо, що він може бути одним з безлічі можливих моделей, які будуть застосовуватися, тоді можна використовувати метод, описаний у [5–8]. Відповідно до цього методу, впорядковані вхідні спостереження $\mathbf{X}s=(x_{(1)}, x_{(2)},.., x_{(n)})^{\mathrm{T}}$ порівнюються з набором серії J з такої самої кількості так званих зразкових спостережень $\mathbf{X}\mathrm{ref}_1=(xref_{1,1},\ xref_{2,1},..,\ xref_{n,1})^T$, $\mathbf{X}\mathrm{ref}_2=(xref_{1,2},\ xref_{2,2},..,\ xref_{n,2})^T$, **…,** $\mathbf{X}\mathrm{ref}_J=(xref_{1,J},\ xref_{2,J},..,\ xref_{n,J})^T$, що відповідають вибраним функціям розподілу ймовірностей $f_1(x), f_2(x), \ldots, f_J(x)$ (рис. 1).

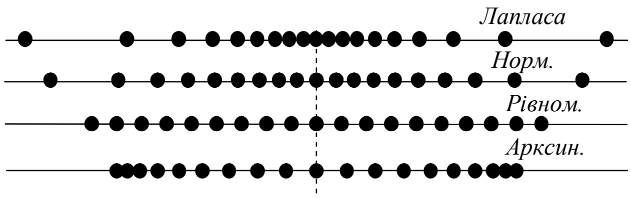

Рис.1. Приклади наборів зразкових спостережень – математичних сподівань позиційних статистик для різних густин розподілів (n=19)

Принцип визначення найкращої оцінки параметра положення (центра групування) результатів спостереження μ і параметра ширини розподілу σ вибірки за допомогою цього методу полягає в попередньому впорядкуванні спостережень $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, а потім у мінімізації суми квадратів S_R^2 відхилень

$$v_k = m + E[s_k]\cdot s - x_{(k)} = xref'_k - x_{(k)},$$

де $xref'_k = m + xref_k \cdot s$ (рис. 2, а).

Для кожної моделі $p_j(x)$ ($j = 1, 2, \ldots, J$) густини розподілу генеральної сукупності розподілу ймовірностей параметри μ_j і σ_j вибірки визначаються на підставі вагового методу найменших квадратів (ВМНК), що у матричному представленні має вигляд [5–8] :

$$(m_j, s_j)^T = \left(\mathbf{A}_j^T \cdot \mathbf{W}_j \cdot \mathbf{A}_j\right)^{-1} \mathbf{A}_j^T \cdot \mathbf{W}_j \cdot \mathbf{X}_s = \mathbf{REC}_j \cdot X_s, \quad (2)$$

де $\mathbf{REC}_j = \left(\mathbf{A}_j^T \cdot \mathbf{W}_j \cdot \mathbf{A}_j\right)^{-1} \mathbf{A}_j^T \cdot \mathbf{W}_j$ (рис. 2, б) є так звана реконструктивна матриця;

$$\mathbf{A}_j^T = \begin{pmatrix} 1 & 1 & \mathbf{L} & 1 \\ xref_{j,1} & xref_{j,2} & \mathbf{L} & xref_{j,n} \end{pmatrix}, \quad \mathbf{W}_j = \mathbf{COV}_j^{-1}$$

– є вагова матриця, зворотна до коваріаційної матриці $\mathbf{COV_j}$ порядкових статистик, елементи якої визначають з обчислення подвійного інтеграла:

$$Cov_{j;k,l} = \iint\limits_{x_l > x_k} s \cdot z \cdot p2_{j;k,l}(s,z)\,ds\,dz - xref_{j;k}\, xref_{j;l}, \quad (3)$$

де

$$p2_{j;k,l}(s,z) = C(n,k,l)\cdot$$
$$\cdot [F_j(s)]^{k-1}[F_j(z)-F_j(s)]^{l-k-1}[1-F_j(z)]^{n-l} p_j(s) p_j(z), \quad (4)$$

є сумісним розподілом ймовірностей k-ї (s) і l-ї (z) порядкових статистик [9]; $F_j(x)$ – функція розподілу;

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

На основі аналізу суми квадратів відхилень

$$S_{R,1}^2, S_{R,2}^2, \ldots, S_{R,j}^2, \ldots, S_{R,J}^2$$

$$S_{R,j}^2 = \frac{\left(\mathbf{X}_s - \mathbf{A}_j\cdot(\hat{m}_j, \hat{s}_j)^T\right)^T \cdot \mathbf{W}_j \cdot \left(\mathbf{X}_s - \mathbf{A}_j\cdot(\hat{m}_j, \hat{s}_j)^T\right)}{n-2}, \quad (5)$$

розраховуються значення для положення і ширини $(\hat{m}; \hat{s})$ вхідної (досліджуваної) вибірки, для яких спостереження в найкращий спосіб (за ВМНК) узгоджуються з відповідною зразковою вибіркою (рис. 2, б).

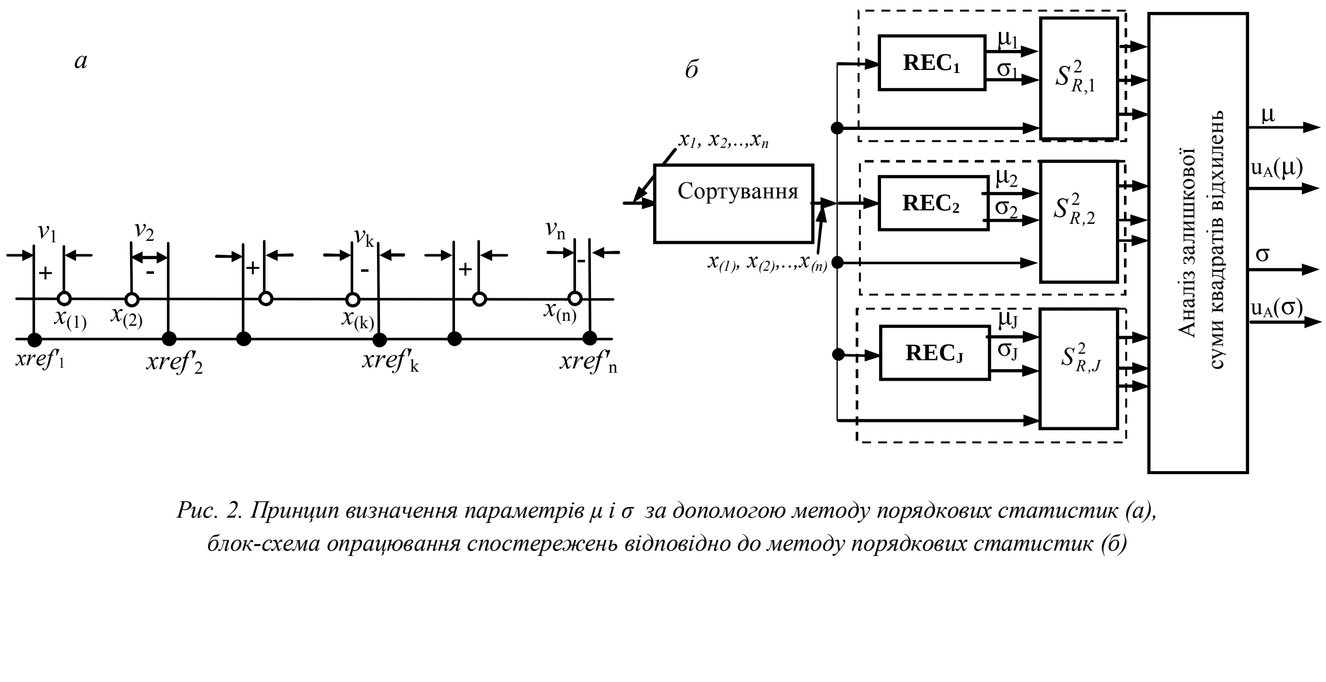

Рис. 2. Принцип визначення параметрів μ і σ за допомогою методу порядкових статистик (а), блок-схема опрацювання спостережень відповідно до методу порядкових статистик (б)

Основною проблемою практичного застосування вищевказаного методу є складність розрахунку коваріаційної матриці **COV.** За виразом (3) для обчислення елемента коваріаційної матриці $Cov_{j;k,l}$ необхідно обчислювати подвійний інтеграл від виразу, що залежить від сумісного розподілу ймовірностей $p2_{j;k,l}(s,z)$ k-ї (s) і l-ї (z) порядкових статистик, який загалом є складною функцією густини та функції розподілу випадкової величини. Крім того, точність розрахунку коваріаційної матриці, а далі оберненої до неї, зменшується зі збільшенням кількості спостережень n.

Метою роботи є розроблення методики опрацювання результатів спостережень на основі наближеного методу порядкових статистик, а також дослідження ефективності методики методом Монте-Карло.

3. Наближений метод порядкових статистик. Для того, щоб значно спростити розрахунок коваріаційної матриці, запропоновано використовувати асимптотичне наближення для дисперсії та коефіцієнтів кореляції між двома позиційними статистиками для більше ніж $n \to \infty$ [9]. Для виведення цих формул використаємо властивості параметрів позиційних статистик, а саме квантилі $x_{(k1)}$ і $x_{(k2)}$ з вибірки, взятої з генеральної сукупності з розподілом $p(x)$, якщо $n \to \infty$, мають асимптотично нормальний розподіл з параметрами [9]:

$$m_1 = x_{(l,1)},\ m_2 = x_{(l,2)}, s_1^2 \approx \frac{l_1(1-l_1)}{n\left(p\left(x_{(l_1)}\right)\right)^2},$$

$$s_2^2 \approx \frac{l_2(1-l_2)}{n\left(p\left(x_{(l_2)}\right)\right)^2},\ r_{1,2} \approx \sqrt{\frac{l_1(1-l_2)}{l_2(1-l_1)}}, \qquad (6)$$

де m_1 m_2 – математичні сподівання (очікувані значення); s_1^2, s_2^2 – дисперсії та $r_{1,2}$ – коефіцієнт кореляції обох квантилів.

Визначаючи значення квантиля для $1 \le k \le n$, як $l_k = k/(n+1)$, для якого $xref_{j;k} \approx x_{(l,j;k)} = qF_j(l_k) = F_j^{-1}(l_k)$ на підставі залежностей (6) наближені значення коефіцієнтів коваріаційної матриці можна обчислити відповідно до співвідношення:

$$Cov_{j;k,l} \approx r_{k,l} \cdot s_{j,k} \cdot s_{j,l} =$$

$$= \frac{k\cdot(n+1-l)}{n(n+1)^2}\cdot\frac{1}{p_j\left(x_{(l,j;k)}\right)\cdot p_j\left(x_{(l,j;l)}\right)},\ 1 \le k < l \le n. \quad (7)$$

4. Дослідження наближеного методу порядкових статистик. Дослідження проводилося у два етапи. У першому порівнювались очікувані значення порядкових статистик і значення елементів автоковаріаційної матриці, які розраховані на основі точних і наближених залежностей, а також значення коефіцієнтів точної та наближеної дворядкової реконструктивної матриці **REC**. Дослідження виконано для кількості спостережень $n = 21$, 41 і 61 і для вибраних ймовірнісних розподілів: Лапласа, нормального і типу арксинусоїдного. Рівномірний розподіл було розглянуто на другому етапі. Виявилося, що значення матриці REC коефіцієнтів, розраховані на основі наближеної коваріаційної матриці (7), менше відрізняються від теоретичних значень, ніж розрахунок на основі «точної» коваріаційної матриці (3).

Зокрема, досліджено вплив наближення на точність розрахунку матриці реконструкції REC на основі наближеної коваріаційної матриці. Для нормального розподілу значення всіх коефіцієнтів першого рядка

$$\mathbf{REC}_j = \begin{pmatrix} g_{1,j} & g_{2,j} & \mathbf{K} & g_{[(n+1)/2],j} & \mathbf{K} & g_{2,j} & g_{1,j} \\ -g_{1,j} & -g_{2,j} & \mathbf{L} & 0 & \mathbf{L} & g_{2,j} & g_{1,j} \end{pmatrix}, \quad (8)$$

мають бути однаковими і дорівнювати $1/n$. Однак у разі обчислення у середовищі Mathcad 13 із точністю представлення чисел 10^{-13}, якщо кількість спостережень n=41 і n=61, значення частини коефіцієнтів істотно відрізняються від $1/n$, причому тим більше, чим більше n (рис. 3). Подібна ситуація спостерігається у випадку інших густин розподілу спостережень. Оскільки неточність матриці **REC** безпосередньо впливає на неточність обчислюваних у (2) параметрів $(\hat{m}; \hat{s})$, то цей факт має важливе значення. З досліджень випливає, що у разі збільшення кількості спостережень точність обчислюваних значень вища в наближеному методі.

На другому етапі ефективність наближеного методу протестовано за допомогою методу Монте-Карло з кількістю симуляцій (реалізацій) M = 10^5. Наведені нижче результати дослідження стосуються значення параметрів положення $\mu_0 = 5{,}000$ і ширини $\sigma = 0{,}200$. Ефективність цих двох методів порівнювали за допомогою статистичного аналізу значень похибок $\Delta_j^{(n)} = \hat{m}_j^{(n)} - m_0$ параметра положення і його стандартної непевності $u_{A,j}^{(n)}(\hat{m})$. Нормалізоване до теоретичного значення стандартного відхилення $s_{\bar{m}} = s \big/ \sqrt{n}$ від середнього значення параметрів похибок і стандартна непевність результатів, отриманих для точних і наближених методів, залежно від кількості спостережень n, подані в таблиці.

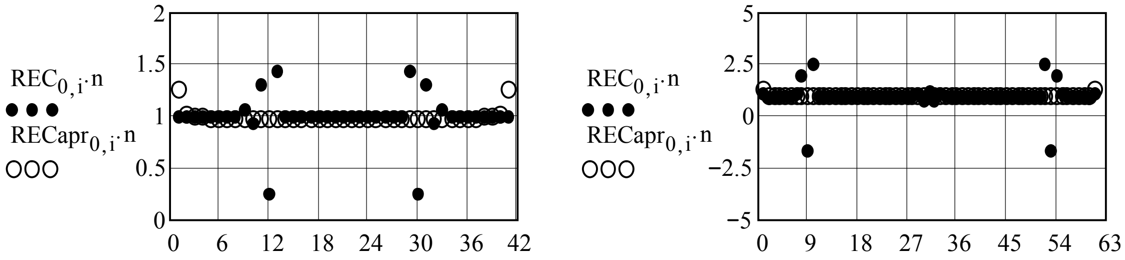

Рис.3. Значення коефіцієнтів першого рядка «точної» матриці REC (●) та наближеної RECapr (○) для нормального розподілу спостережень за кількості n=41 та n=61

Характеристики стандартних похибок і непевності точних і наближених методів

Розподіл Лапласа								
n	Характеристики похибок				Характеристики непевності			
	Точна		Наближена		Точна		Наближена	
	$E_{\bar{\Delta}_{\hat{m}}}/s_{\bar{m}}$	$E_{s_{\Delta(\hat{m})}}/s_{\bar{m}}$	$E_{\bar{\Delta}_{\hat{m}}}/s_{\bar{m}}$	$E_{s_{\Delta(\hat{m})}}/s_{\bar{m}}$	$E_{\bar{u}(\hat{m})}/s_{\bar{m}}$	$E_{s_{u(\hat{m})}}/s_{\bar{m}}$	$E_{\bar{u}(\hat{m})}/s_{\bar{m}}$	$E_{s_{u(\hat{m})}}/s_{\bar{m}}$
21	$3.43\cdot10^{-3}$	0.886	$3.73\cdot10^{-3}$	0.922	0.782	0.247	0.817	0.303
41	$-2.96\cdot10^{-3}$	0.842	$-5.05\cdot10^{-3}$	0.863	0.799	0.193	0.820	0.210
61	$6.77\cdot10^{-3}$	0.831	$8.59\cdot10^{-3}$	0.845	0.800	0.168	0.802	0.180
Розподіл нормальний								
n	Характеристики похибок				Характеристики непевності			
	Точна		Наближена		Точна		Наближена	
	$E_{\bar{\Delta}_{\hat{m}}}/s_{\bar{m}}$	$E_{s_{\Delta(\hat{m})}}/s_{\bar{m}}$	$E_{\bar{\Delta}_{\hat{m}}}/s_{\bar{m}}$	$E_{s_{\Delta(\hat{m})}}/s_{\bar{m}}$	$E_{\bar{u}(\hat{m})}/s_{\bar{m}}$	$E_{s_{u(\hat{m})}}/s_{\bar{m}}$	$E_{\bar{u}(\hat{m})}/s_{\bar{m}}$	$E_{s_{u(\hat{m})}}/s_{\bar{m}}$
21	$4.71\cdot10^{-3}$	1.079	$5.14\cdot10^{-3}$	1.077	0.822	0.266	0.860	0.285
41	$3.25\cdot10^{-3}$	1.030	$3.62\cdot10^{-3}$	1.033	0.937	0.203	0.962	0.211
61	$10.9\cdot10^{-3}$	1.019	$11.0\cdot10^{-3}$	1.019	0.976	0.158	0.993	0.162
Розподіл арксинусоїдний								
n	Характеристики похибок				Характеристики непевності			
	Точна		Наближена		Точна		Наближена	
	$E_{\bar{\Delta}_{\hat{m}}}/s_{\bar{m}}$	$E_{s_{\Delta(\hat{m})}}/s_{\bar{m}}$	$E_{\bar{\Delta}_{\hat{m}}}/s_{\bar{m}}$	$E_{s_{\Delta(\hat{m})}}/s_{\bar{m}}$	$E_{\bar{u}(\hat{m})}/s_{\bar{m}}$	$E_{s_{u(\hat{m})}}/s_{\bar{m}}$	$E_{\bar{u}(\hat{m})}/s_{\bar{m}}$	$E_{s_{u(\hat{m})}}/s_{\bar{m}}$
21	$-8.79\cdot10^{-3}$	0.470	$-8.10\cdot10^{-3}$	0.423	0.269	0.204	0.235	0.259
41	$-2.54\cdot10^{-3}$	0.459	$-2.85\cdot10^{-3}$	0.476	0.168	0.152	0.176	0.175
61	$-0.10\cdot10^{-3}$	0.099	$-0.41\cdot10^{-3}$	0.100	0.134	0.127	0.150	0.141

Розподіл рівномірний				
n	Характеристики непевності (р. рівномірного)			
	$E_{\bar{\Delta}_{\hat{m}}}/s_{\bar{m}}$	$E_{s_{\Delta(\hat{m})}}/s_{\bar{m}}$	$E_{\bar{u}(\hat{m})}/s_{\bar{m}}$	$E_{s_{u(\hat{m})}}/s_{\bar{m}}$
21	$8.93\cdot10^{-3}$	0.771	0.516	0.257
41	$4.12\cdot10^{-3}$	0.619	0.460	0.237
61	$10.3\cdot10^{-3}$	0.552	0.399	0.240

5. Висновки. Як видно з даних, наведених у таблиці, у разі відхилення розподілу ймовірностей нормального розподілу обидва методи дають менші стандартне відхилення похибок і стандартну непевність результату (параметр положення) порівняно зі стандартним відхиленням похибки та стандартною непевністю середнього значення. Зі збільшенням кількості спостережень ефективність точних і наближених методів зростає, причому тим більше, чим більше параметр положення відрізняється від середнього значення.

Різниця між середніми значеннями стандартних відхилень похибок визначає параметр положення в обох методах і не перевищує декількох відсотків. Вона становить:

- для розподілу Лапласа: 4.0 % (*n*=21), 2.5 % (*n*=41), 1.7 % (*n*=61);
- для нормального: 0.26 % (*n*=21), 0.25 % (*n*=41), 0.02 % (*n*=61);
- для типу арксинусоїдного: 9.9 % (*n*=21), 3.4 % (*n*=41), 0.4 % (*n*=61).

Різниця між середніми значеннями стандартної непевності параметра положення в обох методах трохи більша, а саме:

- для розподілу Лапласа: 4,5 % (*n*=21), 2.6 % (*n*=41), 0.3 % (*n*=61);
- для нормального: 4. 6 % (*n*=21), 2.7 % (*n*=41), 1.7 % (*n*=61);
- для типу арксинусоїдного: 12,5 % (*n*=21), 4.9 % (*n*=41), 11.7 % (*n*=61).

Ці дані підтверджують ефективність запропонованого наближеного методу порядкових статистик.
‘

1. Guide of the Expression of Uncertainty in Measurement. International Organisation for Standardisation. – Switzerland, 1993, 1995. 2007. – S. 1–13. 2. Lloyd E.H. Least-squares estimation on location and scale parameters using order statistics. Biometrika, 39 (1952). 88. 3. Downton F. A note of ordered least-squares estimation // Biometrika, 40 (1953). – 457. 4. Kendall M. G. and Stuart A. The Advanced Theory of Statistics, Vol. 2. Charles Griffin and Co Ltd, London. – 3-d edition, 1973. 5. Dorozhovets M. Investigation of the Test Samples Method, Used for the Evaluation of Measurement Result and its Uncertainty. Proc. of Int. Conf. on Precision Measurement. TU Ilmenau. 08–12 Sept. 2008. 91-92. 6. Дорожовець М. Дослідження застосування зразкових вибірок для оцінювання результату вимірювання а його стандартної непевності // Відбір і обробка інформ. – 2008. – Вип. 28 (104). 7. Dorozhovets M. Metoda opracowania wyników obserwacji bazująca na ich porównaniu z prЮбami referencyjnymi // Pomiary. Automatyka. Kontrola. 55 (2009), nr.9. – S. 754–757. 8. Dorozhovets M., Kochan O. Estimation of the best measurement result and its standard uncertainty by input observations processing using the method of reference samples based on order statistics // Proc. of the 5-th IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications. 21–23 September 2009, Rende (Cosenza), Italy. – 351–354. 9. Fisz M. Probability Theory and Mathematical Statistics. – John Willey & Sons, London, 1963.