УДК 519.65, 51-72 doi: 10.20998/2078-5364.2023.3.07

Потаніна Т.В.[1], к.техн.н., доцент, Михайленко І.В.[2], к.пед.н., старш. викладач

ДОСЛІДЖЕННЯ ВИБІРОК ЕКСПЕРИМЕНТАЛЬНИХ ДАНИХ НА НАЯВНІСТЬ ВИКИДІВ: ПОРІВНЯННЯ МЕТОДІВ

[1]Національний технічний університет "Харківський політехнічний інститут", Харків
[2]Харківський національний автомобільно-дорожній університет, Харків

Ключові слова: промахи вимірювань (викиди), мала вибірка, нестатистичні похибки вимірювань, невизначеність, 2σ–критерій, 3σ–критерій, параметричні та непараметричні критерії виявлення викидів, інтервальний аналіз, узгоджена та неузгоджена вибірки, інтервальна статистика.

Постановка проблеми та аналіз публікацій. Виявлення аномальних значень серед результатів спостережень при обробці експериментальних даних і в випадку сукупностей (вибірок) достатньо великого обсягу, а також і для малих вибірок, з великою кількістю аномальних точок і з лише одним викидом, є об'єктом уваги і досліджень науковців в різних царинах протягом вже не одного століття [1–8]. Промахи вимірювань є однією з причин наявності некорисної і шумової інформації в результатах досліджень, а надто велика кількість викидів заважає отримати корисну інформацію і зробити правильні висновки і оцінки процесів, властивостей. Навіщо виявляти такі результати і *з'ясовувати причини* їх наявності? Відповідь здається є очевидною. Конкретизуємо. Викид може свідчити про неправильність результатів спостереження, тобто можливе неправильне кодування даних, некоректне проведення експерименту тощо. З іншого боку, наявність викидів може вказувати не лише на похибки, але й на певні, цікаві з наукової точки зору, факти в проведеному експерименті або на випадкові варіації [9]. При виявленні викидів завжди існує ймовірність допущення похибок 1-го або 2-го роду. Визначення грубих викидів – складна задача, і немає однозначної точки зору щодо її розв'язання.

Розглянемо застосування і ефективність деяких параметричних і непараметричних статистичних критеріїв і підхід інтервальної статистики [8,10] до цієї проблеми.

Викладення основного матеріалу. Найпопулярніша і класична дефініція викиду: викид (промах вимірювання, аномальне значення) – результат спостережень, який найбільше відхилений від середнього значення й не належить генеральному нормальному розподілу випадкової величини. Ключовими словами тут є: нормальний розподіл, середнє та відстань (відхилення).

Зокрема, відстань результату виміру x_i від середнього значення $\overline{x}$ є нормоване відхилення

$$\zeta_i = \left(x_i - \overline{x}\right) \Big/ s_x, i = \overline{1, N}, \tag{1}$$

де стандартне відхилення s_x обчислюється для усіх N точок сукупності

$$s_x = \sqrt{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2 \Big/ \left(N-1\right)}. \qquad (2)$$

Спостереження x_k, для якого $|\zeta_k|$ набуває максимального значення і є викидом.

І вже в цьому означенні є припущення про нормальний розподіл сукупності X, а також застосовується один з його параметрів – середнє значення. Але в більшості випадків ми можемо бути впевнені лише у тому, що дані експерименту *обмежені*. Проте чи сумарна похибка вимірювань є імовірнісною, чи підпорядковується нормальному розподілу, чи похибки послідовних вимірювань незалежні, чи відсутні

Традиційні моделі виявлення викидів, що базуються на понятті відстані (відхилення), можна поділити на декілька класів. Один з них – класичний метод *k-means* (к-середніх) [11], який пропонує кластерізацію даних, застосовуючи класичну евклідову метрику відстані для метрики неподібності. Дослідження показали, що така метрика є досить чутливою до викидів і може бути вдосконалена заміною на алгоритм *к-медіан* і з іншою метрикою неподібності – нормою L_1, або метрикою Манхеттена [12]. Певною мірою такий перехід до іншої метрики і використання медіан замість середніх є спробою перейти від параметричного до непараметричного критерію. Тобто класична оцінка відхилення (1)–(2) також ставиться під сумнів.

Питання впливу обсягу експериментального матеріалу на вибір критеріїв виявлення аномальних значень є надважливим. В випадку сукупностей величезних обсягів задача визначення аномальних значень є досить складною через ресурсний чинник, тому останнім часом для таких вибірок застосовують робастні методи: викиди залишаються у сукупності, однак їхній вплив максимально нівелюється. З іншого боку, чи не одним з найчастіших питань дискусій є проблема вибірки малого розміру, її статистичні оцінки, коректність застосування існуючих непараметричних тестів в тому чи іншому випадку. Роботи, в яких обговорюються і пропонуються методи і підходи до визначення оптимального обсягу вибірки, щоб вона була репрезентативною [13], містять суперечливі аргументи і доведення, а часом і помилкові [14,15]. Разом з тим залишається ціла низка галузей, в яких з різних причин неможливо збільшити обсяг експериментального матеріалу – медицина, психологія, біофізика, дослідження в ядерній енергетиці і таке інше. Проте залишається необхідність зрозуміти, описати і визначити залежності між чинниками, побудувати прогноз, т. і.

Для вибірки з кількістю спостережень N≤8 застосовується методи хроматографічного відокремлення, зокрема критерій Діксона (Q-критерій) розподілу даних експерименту, а в випадку вибірки з N≤6 критерій Граббса ідентифікує не викиди, як викиди, тобто приводить до похибки першого роду (α-похибка). Як було зазначено вище, гарантувати нормальність розподілу даних і похибок вимірювань в більшості випадків практично неможливо.

В ситуаціях з неможливістю припущення про нормальність вибіркового розподілу, при матеріально-технічних, часових та фінансових обмеженнях при зборі даних, в випадку невипадкової вибірки, вибірки малого обсягу або сукупності з викидами, які не можна видалити, застосовуються непараметричні тести виявлення промахів вимірювань U-тест Манна-Уітні, критерії Вілкоксона, Крускала-Уоліса і т.д. [16–19].

Розглянемо застосування статистичних критеріїв та критеріїв інтервальної статистики для розв’язання задачі виявлення аномальних значень, і порівняємо результати.

Актуальною задачею в дослідженні властивостей металів ядерної чистоти є визначення залежності твердості гафнію по Брінеллю від домішок газів. Проведемо обробку вибірки зашумлених експериментальних даних з фіксованим значенням основного аргументу (відсотковий ваговий вміст кисню) та багаторазовими вимірюваннями твердості по Брінеллю злитків гафнію. Результати вимірювань даної характеристики гафнієвих зразків представлені на рисунку 1. Дана вибірка містить 24 спостереження:

$$\left\{x_i, i=\overline{1,24}\right\} = \{178, 176,\ 187,\ 184,\ 180,\ 178,\ 182,\ 185,\ 179,\ 187,\ 182,\ 174, 179,\ 177,\ 180,$$
$$178,\ 178,\ 183,\ 176,\ 198,\ 195,\ 187,\ 196,\ 195\}\ \text{НВ}.$$

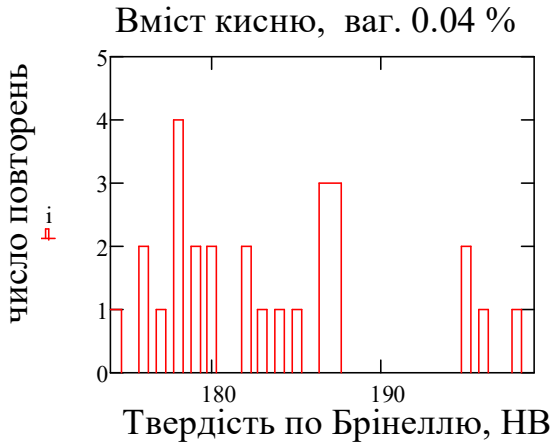

Рисунок 1 – Результати вимірювань твердості по Брінеллю злитків гафнію при вмісті кисню 0,04 ваг. %.

Значення вмісту кисню достовірно відоме, а вимірюванні значення твердості містять як звичайні інструментальні похибки вимірювань, так і хаотичні викривлення невідомої величини. Ймовірнісні характеристики обох компонентів похибки невідомі або не підпорядковуються закону нормального розподілу; обмеження на максимальне значення сумарних похибок також є невизначеним.

Неможливо гарантувати, що спостереження проводяться з однаковою точністю. Тому, згідно зі стандартом [20], стандартна інструментальна похибка вимірювання твердості по Брінеллю приймається рівною 4 %. Це значення допускається, якщо діапазон виміряної твердості по Брінеллю становить (200 ± 50)НВ. Таким чином, обмежуються лише цією похибкою вимірювань і не враховують інші можливі складові сумарної похибки вимірювання [21].

В інтервальній статистиці базовим поняттям є множина невизначеності виміру, яка гарантовано містить істинне значення спостереження. Множини невизначеності мають наступний вигляд:

$$H_i = \left[\underline{h_i}; \overline{h_i}\right] = \left[x_i - \Delta_i; x_i + \Delta_i\right], \tag{3}$$

де Δ_i – абсолютна похибка в i-му спостереженні.

Використовуючи межі множин невизначеності H_i, розраховуються допоміжні екстремальні значення (для усіх спостережень $i = \overline{1, N}$):

$$h_{\min} = \max_i \underline{h_i},\ h_{\max} = \min_i \overline{h_i}\ . \tag{4}$$

Для даної вибірки отримуємо інтервали :

$$H_1 = [170.88;185.12] = H_6 = H_{16} = H_{17}, H_2 = [168.96;183.04] = H_{19},$$
$$H_3 = [179.52;194.48] = H_{10} = H_{22},\ H_4 = [176.64;191.36],\ H_5 = H_{15} = [172.8;187.2],$$
$$H_7 = [174.72;189.28] = H_{11}, H_8 = [177.6;192.4],\ H_9 = [171.84;186.16] = H_{13},$$
$$H_{12} = [167.04;180.96],\ H_{14} = [169.92;184.08],\ H_{18} = [175.68;190.32],$$
$$H_{20} = [190.08;205.92],\ H_{21} = [187.2;202.8] = H_{24},\ H_{23} = [188.16;203.84].$$

Екстремальні значення $h_{\min}, h_{\max}$:

$$h_{\min} = 190.08, h_{\max} = 180.96\ .$$

Очевидно, що $h_{\min} > h_{\max}$ – тому вибірка вважається неузгодженою. Слід провести додатковий аналіз даної сукупності.

Процедура пошуку одиночних викидів:

1) Обчислення множини попарних перетинів усіх множин H_i і H_j :

$$P_{ij} = H_i \cap H_j, i = \overline{1, N-1}, j = \overline{i+1, N}\ , \tag{5}$$

причому межі множини P_{ij} :

$$\underline{p_{ij}} = \max\left\{\underline{h_i}, \underline{h_j}\right\}, \overline{p_{ij}} = \min\left\{\overline{h_i}, \overline{h_j}\right\}. \tag{6}$$

2) Якщо $\underline{p_{ij}} > \overline{p_{ij}}$, тоді $P_{ij} = \varnothing$ і значення атрибута сумісності $S_{ij} = 0$;

Якщо $\underline{p_{ij}} \le \overline{p_{ij}}$, тоді $P_{ij} \neq \varnothing$ і значення атрибута сумісності $S_{ij} = 1$.

3) Побудова таблиці сумісності

$$\left\{S_{ij}, S_{ij} = S_{ji,}\right\}, i = \overline{1, N-1}, j = \overline{i+1, N}\ . \tag{7}$$

4) Номер спостереження i є одиночним викидом і видаляється з вибірки, якщо відповідний рядок у таблиці сумісності складається з нулів. Поодинокі викиди видаляються, а вибірка усікається.

Далі за допомогою таблиці сумісності визначаємо сумісну підвибірку максимальної довжини.

Обирається рядок m (або рядки $m_1,...,m_l$) з максимальною кількістю одиниць, що містяться в стовпцях $j_1,...,j_k$. Обирається перший із стовпців j_1 з одиничним значенням атрибута сумісності. Усі елементи стовпців переглядаються зверху вниз. Елементи з $S_{ij}=0$ вилучаються з послідовності $j_1,...,j_k$. Операція повторюється для всіх стовпців. Отримуємо набір номерів спостережень, які складають сумісну підвибірку максимальної довжини.

Матриця сумісності для упорядкованої вибірки представлено в таблиці 1.

Тривіальні комірки, розташовані на головній діагоналі таблиці, не беруть участі в аналізі сумісності вибірки.

i/j	**1**	**2**	**…**	**20**	**21**	**22**	**23**	**24**
1		1	…	0	0	1	0	0
2	1		…	0	0	1	0	0
…	…	…		…	…	…	…	…
20	0	0	…		1	1	1	1
21	0	0	…	1		1	1	1
22	1	1	…	1	1		1	1
23	0	0	…	1	1	1		1
24	0	0	…	1	1	1	1	

Таблиця 1 – Атрибути попарної сумісності множин невизначеності

Використовуючи представлений алгоритм, визначимо, що спостереження 20, 21, 22, 23, 24 є викидами. Побудуємо усічену вибірку. Для даної вибірки:

$$h_{\min}=179.52,\ h_{\max}=180.96,$$

тобто $h_{\min}\le h_{\max}$ і вибірка сумісна.

Інформаційна множина: $I=[179.52,180.96]$.

Оцінка центрального фактичної значення

$$x_c=\frac{h_{\min}+h_{\max}}{2},\quad x_c=180.24. \tag{8}$$

Максимальне фактичне відхилення

$$\Delta x=\frac{h_{\max}-h_{\min}}{2},\qquad \Delta x=0.72. \tag{9}$$

Застосуємо стандартні статистичні методи для обробки вибірки. Такий підхід є формальним, оскільки ймовірнісні характеристики похибки спостереження невідомі.

Середнє вибіркове

$$\bar{x} = \frac{\sum_i x_i}{N} = 183.083. \tag{10}$$

Стандартне відхилення

$$\sigma = \sqrt{\left[\sum_i \left(x_i - \bar{x}\right)^2\right] \Big/ \left[N-1\right]} = 6,921. \tag{11}$$

Вибіркове середнє не належить до інформаційної множини. Стандартне відхилення майже у 96 разів перевищує максимальне фактичне відхилення.

Формальне застосування 3σ-критерія для виявлення викидів приводить до висновку, що вибірка не містить значущих викидів: усі значення належать інтервалу

$$\left[\bar{x} - 3\sigma; \bar{x} + 3\sigma\right] = \left[162.32; 203.85\right].$$

Критерій 2σ дозволяє зробити висновок, що експеримент номер три є викидом:

$$198 \notin \left[\bar{x} - 2\sigma; \bar{x} + 2\sigma\right] = \left[169.24; 196.25\right].$$

Інший спосіб визначення викидів — це використання медіани вибірки. Медіана – проста й найбільш надійна оцінка параметра зсуву для вибірки з невеликою кількістю аномальних даних. Для таких даних вибіркове середнє може дати незадовільний результат. Викидами можна вважати точки, що знаходяться за межами

$$\left[\text{med} - 3 \cdot \sigma\left(\text{med}\right); \text{med} + 3 \cdot \sigma\left(\text{med}\right)\right], \tag{12}$$

де med – медіана, $\sigma\left(\text{med}\right)$ – середня квадратична похибка для медіани. Застосування 3σ-критерію для медіани вибірки $\text{med} = 181$ та середньоквадратичної похибки для медіани $\sigma\left(\text{med}\right) = 7.241$ не виявляє викидів у вибірці. Інтервал (12) містить всі точки вибірки: $\left[159.28; 202.72\right]$.

Згідно 2σ-критерію, викидом є значення 198 НВ; за 3σ-критерієм викидів немає; 3σ-криитерій застосований для медіани також не визначає викидів.

Критерій Львовського виявлення викидів є параметричним критерієм і передбачає, що вибірка має нормальний розподіл і середньою за обсягом. «Підозріле» спостереження вважається викидом і виключається з вибірки, якщо для табличного значення критерію Львовського KrL вірне співвідношення:

$$KrL < \frac{\left|\tilde{x} - \bar{x}\right|}{\sqrt{D} \cdot \sqrt{n-1/n}}, \tag{13}$$

де D – дисперсія.

Обчислимо критерій Львовського для «підозрілих» значень 195, 196, 198 НВ (13): табличне значення $KrL(24) = 2.7$, і

$$2.7 > \frac{|195-183.083|}{6.776\cdot\sqrt{23/24}} = 1.797,\ \ 2.7 > \frac{|196-183.083|}{6.776\cdot\sqrt{23/24}} = 1.947,\ 2.7 > \frac{|198-183.083|}{6.776\cdot\sqrt{23/24}} = 2.249.$$

Таким чином, дані точки не є викидами згідно критерію Львовського.

Одним з поширених методів аналізу викидів є правило «скриньки з вусами» (box-and-whiskers-plot), який послугується поняттями медіани вибірки, Q_1 і Q_3 квартилями, міжквартильним розмахом $R = Q_3 - Q_1$.

Очевидно, що для вибірки значень твердості зразків гафнію в випадку, що досліджується:

$$Q_1 = 178,\ Q_3 = 187,\ \ R = 9.$$

Відповідно, до м'яких (підозрілих) викидів належатимуть точки 195,196,198, тому що вони знаходяться поза інтервалом

$$\left[\text{med} - 1.5\cdot R; \text{med} + 1.5\cdot R\right] = [167.23; 194.77];$$

Екстремальних викидів, які обов'язково мають бути видаленими з вибірки, немає, бо точок поза проміжком

$$\left[\text{med} - 3\cdot R; \text{med} + 3\cdot R\right] = [154; 208]$$

не виявлено.

Даний критерій підтверджує результати інтервальної статистики.

Висновки. Використання методів інтервального аналізу є альтернативним гнучким інструментарієм для отримання більш точного та повного аналізу експериментальних даних за наявності неповної інформації, шумів, викидів вимірювань, що характерно при проведенні досліджень процесів, властивостей матеріалів, систем в різноманітних галузях. Підхід інтервальної статистики може бути використаний для підтвердження результатів застосування різних критеріїв або в таких ситуаціях, коли їх застосування некоректне. Виявлення викидів методами інтервального аналізу не є чутливим до обсягу вибірки, до кількості аномальних значень, до того, чи розподіл спостерігаємої величини є нормальним.

Література

1. Wada K. Outliers in official statistics // Japanese Journal of Statistics and Data Science. 2020. No 3. pp. 669–691. https://doi.org/10.1007/s42081-020-00091-y.

2. Barnett V., Lewis, T. Outliers in statistical data. John Wiley & Sons Chichester, West Sussex. 1994. 584 p.

3. Barnett V. Outliers in sample surveys // Journal of applied statistics. 1994. No 21. pp. 381–389.

4. Orellana M., Cedillo P. Outlier Detection with Data Mining Techniques and Statistical Methods // InternationalConference on Information Systems and Computer Science (INCISCOS). 2019. https://doi.org/10.1109/INCISCOS49368.2019.00017.

5. Peirce B. Criterion for the rejection of doubtful observations // Astronomical Jour-

nal II. 1852. No 45. pp. 161–163.

6. Bertarelli G., Chambers R., Salvati N. Outlier robust small domain estimation via bias correction and robust bootstrapping // Statistical Methods and Applications, Springer, Societa Italiana di Statistica. 2021. Vol. 30(1). pp. 331–357. http://doi.org/10.1007/s10260-020-00514-w.

7. Лихач О.Ю., Угрюмов М.Л., Шевченко Д.О., Шматков С.І. Методи виявлення викидів в пробних вибірках при управлінні процесами в системах за станом // Вісник Харківського національного університету імені В.Н. Каразина, серія «Математичне моделювання. Інформаційні технології. Автоматизовані системи управління». 2022. Вип. 53. С.2 1–40. https://doi.org/10.26565/2304-6201-2022-53-03.

8. Yefimov O.V., Potanina T.V. Determination of the dependence of the NPP unit power on the steam temperature at the outlet of the superheater separator first stage with an uncertainty of information // Problems of Atomic Science and Technology. 2022. No. 137(1). pp. 169–172.

9. NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/,date. 2012.

10. Moore R.E., Kearfott R.B., Cloud M.J. Introduction to interval analysis. Philadelphia. Society for Industrial and Applied Mathematics, 2009. 223 p.

11. Na S., Xumin L., Yong G. Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm // Third International Symposium on Intelligent Information Technology and Security Informatics. 2010. IEEE. Jian, China. pp. 63–67. doi: 10.1109/IITSI.2010.74.

12. Helm M. Use this clustering method if you have many outliers. The k-medians variation for robust outcomes // Towards Data Science. 2021.
https://towardsdatascience.com/use-this-clustering-method-if-you-have-many-outliers-5c99b4cd380d

13. Олефір В., Боснюк В. Розрахунок обсягу вибірки як наріжний камінь планування наукового дослідження // Вісник Львівського Університету. Серія Психологічні науки. 2021. № 9. С. 186–195. doi:10.30970/PS.2021.9.24.

14. Horton R. Offline: What is medicine's 5 sigma? // The Lancet. 2015. Vol. 385, Issue 9976. p. 1380. doi:10.1016/S0140-6736(15)60696-1.

15. Reiczigel J., Rozsa L. Do small samples underestimate mean abundance? It depends on what type of bias we consider // Folia Parasitologica. 2017. No. 64:025. doi:10.14411/fp.2017.025.

16. Hollander Myles, Wolfe Douglas A., Chicken E. Nonparametric Statistical Methods // John Wiley & Sons. 2014. Inc. ISBN 978-0-470-38737-5. pp. 39–41.

17. Zhexue H. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values // Data Mining and Knowledge Discovery. Kluwer Academic Publishers. Manufactured in The Netherlands.1998. No. 2. pp. 283–304.

18. Patrick A. Regoniel. Nonparametric Tests: 8 Important Considerations in Using Them // Research-based Articles. 2020. https://simplyeducate.me/2020/10/11/nonparametric-tests/.

19. Winter J.C.F., Dodou D. Five-Point Likert Items: t test versus Mann-Whitney-Wilcoxon // Practical Assessment. Research and Evaluation. 2010. No. 15(11).

20. Національний Стандарт України. Матеріали металеві. Випробування на твердість по Брінеллю. Частина 2. Перевірення та калібрування випробувальних машин (ДСТУ EN ISO 6506-2:2019, IDT). Київ, 2019. 25 с.

21. Настанова державного підприємства «НАЕК «Енергоатом»»: Методика ви-

значення механічних властивостей металу за результатами випробувань на твердість. Київ, 2016, 33 с.

Bibliography (transliterated)

1. Wada K. Outliers in official statistics // Japanese Journal of Statistics and Data Science. 2020. No 3. pp. 669–691. https://doi.org/10.1007/s42081-020-00091-y.

2. Barnett V., Lewis, T. Outliers in statistical data. John Wiley & Sons Chichester, West Sussex. 1994. 584 p.

3. Barnett V. Outliers in sample surveys // Journal of applied statistics. 1994. No 21. pp. 381–389.

4. Orellana M., Cedillo P. Outlier Detection with Data Mining Techniques and Statistical Methods // InternationalConference on Information Systems and Computer Science (INCISCOS). 2019. https://doi.org/10.1109/INCISCOS49368.2019.00017.

5. Peirce B. Criterion for the rejection of doubtful observations // Astronomical Journal II. 1852. No 45. pp. 161–163.

6. Bertarelli G., Chambers R., Salvati N. Outlier robust small domain estimation via bias correction and robust bootstrapping // Statistical Methods and Applications, Springer, Societa Italiana di Statistica/ 2021. Vol. 30(1). pp. 331–357. http://doi.org/10.1007/s10260-020-00514-w.

7. Lykhach O.Yu., Ugryumov M.L., Shevchenko D.O., Shmatkov S.I. Metody vyiavlennia vykydiv v probnykh vybirkakh pry upravlinni protsesamy v systemakh za stanom // Visnyk Kharkivskoho natsionalnoho universytetu imeni V.N. Karazyna, seriia «Matematychne modeliuvannia. Informatsiini tekhnolohii. Avtomatyzovani systemy upravlinnia». 2022. No. 53. pp.21- 40. https://doi.org/10.26565/2304-6201-2022-53-03.

8. Yefimov O.V., Potanina T.V. Determination of the dependence of the NPP unit power on the steam temperature at the outlet of the superheater separator first stage with an uncertainty of information // Problems of Atomic Science and Technology. 2022. No. 137(1). pp. 169–172.

9. NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/,date. 2012.

10. Moore R.E., Kearfott R.B., Cloud M.J. Introduction to interval analysis. Philadelphia. Society for Industrial and Applied Mathematics, 2009. 223 p.

11. Na S., Xumin L., Yong G. Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm // Third International Symposium on Intelligent Information Technology and Security Informatics. 2010. Jian, China. pp. 63–67. doi: 10.1109/IITSI.2010.74.

12. Helm M. Use this clustering method if you have many outliers. The k-medians variation for robust outcomes // Towards Data Science. 2021.

https://towardsdatascience.com/use-this-clustering-method-if-you-have-many-outliers-5c99b4cd380d

13. Olefir V., Bosniuk V. Rozrakhunok obsiahu vybirky yak narizhnyi kamin planuvannia naukovoho doslidzhennia // Visnyk Lvivskoho Universytetu. Seriia Psykholohichni nauky. 2021. Issue 9. pp. 186–195. doi:10.30970/PS.2021.9.24.

14. Horton R. Offline: What is medicine's 5 sigma? // The Lancet. 2015. Vol. 385, Issue 9976. p. 1380. doi:10.1016/S0140-6736(15)60696-1.

15. Reiczigel J., Rozsa L. Do small samples underestimate mean abundance? It depends on what type of bias we consider // Folia Parasitologica. 2017. No. 64:025.

doi:10.14411/fp.2017.025.

16. Hollander Myles, Wolfe Douglas A., Chicken E. Nonparametric Statistical Methods // John Wiley & Sons. 2014. Inc. ISBN 978-0-470-38737-5. pp. 39–41.

17. Zhexue H. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values // Data Mining and Knowledge Discovery. Kluwer Academic Publishers. Manufactured in The Netherlands.1998. No. 2. pp. 283–304.

18. Patrick A. Regoniel. Nonparametric Tests: 8 Important Considerations in Using Them // Research-based Articles. 2020. https://simplyeducate.me/2020/10/11/nonparametric-tests/.

19. Winter J.C.F., Dodou D. Five-Point Likert Items: t test versus Mann-Whitney-Wilcoxon // Practical Assessment. Research and Evaluation. 2010. No. 15(11).

20. Natsionalnyi Standart Ukrainy. Materialy metalevi. Vyprobuvannia na tverdist po Brinelliu. Chastyna 2. Perevirennia ta kalibruvannia vyprobuvalnykh mashyn (DSTU EN ISO 6506-2:2019, IDT). Kyiv, 2019. 25 p.

21. Nastanova derzhavnoho pidpryiemstva «NAEK «Enerhoatom»»: Metodyka vyznachennia mekhanichnykh vlastyvostei metalu za rezultatamy vyprobuvan na tverdist. Kyiv, 2016. 33 p.

УДК 519.65, 51-72

Потаніна Т.В., к.техн.н., доцент, Михайленко І.В., к.пед.н., старш. викладач

ДОСЛІДЖЕННЯ ВИБІРОК ЕКСПЕРИМЕНТАЛЬНИХ ДАНИХ НА НАЯВНІСТЬ ВИКИДІВ: ПОРІВНЯННЯ МЕТОДІВ

Задача виявлення викидів (промахів, аномальних значень, результатів, що різко виділяються, результатів, що відірвалися) є однією з найактуальніших, складних і неоднозначних при обробці експериментального матеріалу. Такими значеннями вважаються результати експерименту, які знаходяться аномально далеко від інших точок із серії паралельних спостережень.

Джерелом викидів нерідко є похибки вимірювань. Серед таких є невірний запис результатів експерименту, можливе неправильне кодування даних, некоректне проведення експерименту тощо. Грубі похибки виникають при різкій зміні умов проведення дослідження, несправностях в роботі апаратури й т.і.

Одночасно викиди можуть свідчити про неочікувану, неординарну поведінку вимірюваної величини, яка є проявом ще не з'ясованої властивості процесу. І тому потрібен аналіз з застосуванням надійного математичного інструментарія.

Методи виявлення викидів різноманітні і численні. Параметричні тести мають більшу чутливість до розміру вибірки і до ймовірнісного розподілу значень сукупності. Більш гнучкими є непараметричні тести, які можна застосувати, якщо не можна зробити припущення про нормальність вибіркової сукупності або обсяг вибірки малий; такі критерії дають кращий результат в асиметричних розподілах, тому що застосовують медіану замість середнього; їх можна застосовувати для порядкових або номінальних даних, а також в ситуації аберрантного значення викиду.

Методи інтервального аналізу, зокрема інтервальної статистики, є альтернативним гнучким інструментарієм для отримання більш точного та повного аналізу експериментальних даних за наявності неповної інформації, шумів, викидів вимірювань, наявності аномальних та аберрантних точок.

Проведено порівняння результатів застосування параметричних критеріїв (2σ-критерій, 3σ-критерій, Львовського) та непараметричних критеріїв (правило «скриньки з вусами») виявлення викидів, а також обчислення методами інтервальної статистики. Один з викидів був визначений таким непараметричним критерієм, 3σ-критерієм і процедурою виявлення поодинокого викиду інтервальними методами. Ще два значення були виявлені, як підозрілі викиди за допомогою правила «скринька з вусами» і алгоритму розпізнання з інтервальної статистики.

Методи виявлення викидів методами інтервального аналізу є не менш ефективними, ніж застосування непараметричних тестів.

Ключові слова: промахи вимірювань (викиди), мала вибірка, нестатистичні похибки вимірювань, невизначеність, 2σ-критерій, 3σ-критерій, параметричні та непараметричні критерії виявлення викидів, інтервальний аналіз, узгоджена та неузгоджена вибірки, інтервальна статистика.

Потанина Т.В., к.техн.н., доцент, Михайленко И.В., к.пед.н., старш. преподаватель

ИССЛЕДОВАНИЕ ВЫБОРОК ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ НА НАЛИЧИЕ ВЫБРОСОВ: СРАВНЕНИЕ МЕТОДОВ

Задача обнаружения выбросов (промахов, аномальных значений, резко выделяющихся результатов, оторвавшихся результатов) является одной из наиболее актуальних, сложных и неоднозначных при обработке экспериментального материала. Такими значениями считаются результаты эксперимента, находящиеся аномально далеко от других точек в серии параллельных наблюдений.

Источником выбросов часто являються ошибки и погрешности измерений. Среди таких – неверная запись результатов эксперимента, возможное неправильное кодирование данных, некорректное проведение эксперимента и др. Грубые ошибки возникают при резком изменении условий проведения исследования, неисправностях в работе аппаратуры и т.д.

В тоже время выбросы можуть свидетельствовать о неожиданном, неординарном поведении измеряемой величины, которое является проявлением еще не выясненных свойств процесса. И потому необходим анализ с применением надежного математического инструментария.

Методы обнаружения выбросов разнообразны и многочисленны. Параметрические тесты имеют большую чувствительность к размеру виборки и к вероятностному распределению значений совокупности. Более гибкими являются непараметрические тесты, которые можно применять, если нельзя предположить нормальность выборочной совокупности или объем выборки мал; такие критерии дают лучший результат в ассиметричных распределениях, поскольку используют медиану вместо среднего; их можно применять для порядковых или номинальных данных, а также в ситуации аберрантного значения выброса.

Методы интервального анализа, в частности интервальной статистики – альтернативный гибкий инструментарий для получения более точного и полного анализа экспериментальных данных в случае неполноты информации, шумов, выбросов измерений, наличия аномальных и аберрантных точок.

Проведено сравнение результатов применения параметрических критериев (2σ-критерий, 3σ-критерий, Львовского) и непараметрических критериев (правило «ящика с усами») обнаружения выбросов, а также вычисления методами интервальной статис-

тики. Один из выбросов был выявлен непараметрическим критерием, 3σ-критерием и процедурой обнаружения одиночных выбросов интервальными методами. Еще два значения были определены, как подозрительные выбросы с помощью правила «ящик с усами» и алгоритма распознания из интервальной статистики.

Методы обнаружения выбросов методами интервального анализа являются не менее эффективными, чем проверка непараметрическими тестами.

Ключевые слова: промахи измерений (выбросы), малая выборка, нестатистические ошибки измерений, неопределенность, 2σ–критерий, 3σ–критерий, параметрические и непараметрические критерии, интервальный анализ, согласованная и несогласованная выборки, интервальная статистика.

Potanina T.V., Mykhaylenko I.V.

EXAMINATION OF EXPERIMENTAL DATA SAMPLES FOR THE PRESENCE OF OUTLIERS: COMPARISON OF METHODS

The task of detecting outliers (misses, abnormalous values, results that stand out sharply, results that have come off) is one of the most relevant, complex and ambiguous in the experimental materialprocessing. Such values are the experiment results, which are abnormally far from other points from a series of parallel observations.

The source of emissions is often measurement errors. Among these are incorrect recording of the experiment results, possible incorrect coding of data, incorrect conduct of the experiment, etc. Gross errors occur in the event of a sudden change in the conditions of conducting the research, malfunctions in the operation of the equipment, etc.

At the same time, outliers may indicate an unexpected, extraordinary behavior of the measured value – a yet-to-be-explained property process manifestation. And that's why an analysis using reliable mathematical tools is needed.

The methods of detecting emissions are diverse and numerous. Parametric tests are more sensitive to the sample size and to the population values probability distribution. Non-parametric tests are more flexible and can be applied if the non-normal distributon of the sample or the sample size is small; such criteria give a better result in asymmetric distributions, because they use the median instead of the mean; they can be applied to ordinal or nominal data, as well as in the situation of an aberrant outlier value.

Interval analysis methods, in particular interval statistics, are an alternative flexible toolkit for obtaining a more accurate and complete analysis of experimental data in the incomplete information, noise presence, measurement outliers, and the presence of abnormalous and aberrant points.

A comparison of the results of the application of parametric criteria (2σ-criterion, 3σ-criterion, Lvovskyi) and non-parametric criteria (the box-and-whiskers-plot) for detecting emissions, as well as calculation using interval statistics methods, was carried out. One of the outliers was determined by the non-parametric criterion, the -criterion and the procedure for detecting a single outlier using interval methods. Two values are suspicious outliers using the box-whisker rule and the interval statistics recognition algorithm.

The methods of detecting outliers using interval analysis methods are no less effective than the use of non-parametric tests.

Keywords: outliers, small sample, non statistical measurement errors, uncertainty, two sigmas criterion, three sigmas criterion, parametric and non-parametric detection outliers criterions, interval analysis, compatible and incompatible sample, interval statistics.