

ЗАСТОСУВАННЯ МЕТОДІВ МАТЕМАТИЧНОЇ СТАТИСТИКИ В ПРОГРАМНІЙ ОБРОБЦІ ДАНИХ

Вінницький національний технічний університет

Анотація

У статті розглянуто основні методи математичної статистики та їх застосування в програмній обробці даних. Проаналізовано базові статистичні характеристики, зокрема середнє значення, дисперсію та стандартне відхилення, а також методи кореляційного і регресійного аналізу. Показано значення використання статистичних методів у сучасних програмних системах та інформаційних технологіях.

Ключові слова: математична статистика, програмна обробка даних, аналіз даних, кореляція, регресія.

Abstract

The article considers the basic methods of mathematical statistics and their application in software data processing. Basic statistical characteristics such as mean value, variance, and standard deviation are analyzed, as well as correlation and regression analysis methods. The importance of statistical methods in modern software systems and information technologies is shown.

Keywords: mathematical statistics, data processing, data analysis, correlation, regression.

У сучасному світі інформаційних технологій обсяги даних, які створюються та обробляються програмними засобами, постійно зростають. Такі дані використовуються в різних галузях, зокрема в економіці, медицині, освіті, інженерії та соціальних науках. У зв'язку з цим виникає необхідність ефективного аналізу інформації з метою отримання достовірних результатів і прийняття обґрунтованих рішень.

Важливу роль у процесі аналізу даних відіграють методи математичної статистики. Вони дозволяють узагальнювати вибіркові дані, виявляти закономірності та оцінювати параметри досліджуваних процесів. Поєднання статистичних методів із програмуванням забезпечує автоматизацію обчислень і значно зменшує вплив людського фактора [1].

Основні поняття математичної статистики

Математична статистика в основному працює з вибірковими даними, які є частиною генеральної сукупності. До найпоширеніших статистичних характеристик належать середні значення та показники розсіювання даних [2].

У програмній обробці даних, особливо при наявності шумів та аномальних значень, замість класичного середнього арифметичного часто використовується усічене середнє значення, яке обчислюється за формулою:

$$X_{trim} = \frac{1}{n - 2k} \sum_{i=k+1}^{n-k} x_{(i)}$$

де $x_{(i)}$ - впорядковані за зростанням значення вибірки, k - кількість крайніх елементів, що відкидаються.

Такий підхід дозволяє зменшити вплив викидів і є доцільним у програмному аналізі реальних даних.

Для оцінювання розсіювання даних у програмних системах також застосовується середнє абсолютне відхилення від медіани, яке визначається як:

$$MAD = \frac{1}{n} \sum_{i=1}^n |x_i - med(x)|$$

Цей показник є робастною характеристикою розсіювання та широко використовується в прикладному аналізі даних, оскільки він менш чутливий до аномальних значень.

Для отримання оцінки стандартного відхилення на основі середнього абсолютного відхилення використовується співвідношення:

$$\delta_{rob} = 1.4826 \times MAD$$

Окрім цього, важливе значення мають методи аналізу залежностей між змінними. У випадках, коли зв'язок між даними не є лінійним або дані містять викиди, використовується коефіцієнт рангової кореляції Спірмена, який обчислюється за формулою:

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

де d_i - різниця між рангами відповідних значень змінних.

Для оцінювання якості побудованих моделей у задачах регресійного аналізу використовується середньоквадратична помилка, яка визначається як:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \widehat{y}_i)^2$$

Цей показник дозволяє кількісно оцінити точність прогнозування та широко використовується в програмних реалізаціях статистичних і машинних моделей.

Кореляційний аналіз дозволяє визначити ступінь взаємозв'язку між змінними, тоді як регресійний аналіз використовується для побудови математичних моделей і прогнозування значень [3].

Реалізація статистичних методів у програмуванні

У програмній обробці даних статистичні методи реалізуються за допомогою алгоритмів та структур даних. Сучасні мови програмування, такі як Python, Java та C++, надають широкі можливості для роботи з числовими масивами та великими обсягами інформації.

Прості статистичні показники, наприклад середнє значення або стандартне відхилення, можуть бути реалізовані за допомогою базових циклів і математичних операцій. Для складніших задач, таких як кореляційний аналіз або побудова регресійних моделей, використовуються спеціалізовані бібліотеки, зокрема NumPy, SciPy та pandas [4].

Автоматизація статистичних обчислень у програмному середовищі дозволяє швидко обробляти великі масиви даних, що є особливо важливим при роботі з технологіями Big Data.

Застосування статистичних методів в інформаційних системах

Методи математичної статистики широко застосовуються в аналізі даних та інформаційних системах. Вони використовуються для попередньої обробки інформації, виявлення аномалій, класифікації даних та прогнозування.

У інформаційних системах статистичні методи дозволяють аналізувати поведінку користувачів, оцінювати ефективність роботи програмного забезпечення та оптимізувати його функціонування. Наприклад, аналіз середніх значень і дисперсії дає змогу оцінити навантаження на сервери або стабільність роботи системи [5].

Крім того, статистичні методи є основою для машинного навчання та штучного інтелекту, де вони застосовуються для навчання моделей і оцінювання якості отриманих результатів.

Методи математичної статистики є важливою складовою сучасної програмної обробки даних. Їх використання дозволяє ефективно аналізувати інформацію, виявляти закономірності та приймати обґрунтовані рішення. Поєднання статистичних методів із програмуванням забезпечує автоматизацію обчислень, підвищує точність аналізу та розширює можливості інформаційних систем.

Таким чином, застосування методів математичної статистики в програмній обробці даних залишається актуальним напрямом досліджень і має велике практичне значення.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Montgomery D. C., Runger G. C. *Applied Statistics and Probability for Engineers*. – 7th ed. – Hoboken: John Wiley & Sons, 2018. URL: <https://www.wiley.com/en-us/Applied+Statistics+and+Probability+for+Engineers-p-9781118539712> (дата звернення: 18.12.2025).
2. Walpole R. E., Myers R. H., Myers S. L., Ye K. *Probability and Statistics for Engineers and Scientists*. – 9th ed. – Boston: Pearson Education, 2017. URL: <https://www.pearson.com/en-us/subject-catalog/p/probability-and-statistics-for-engineers-and-scientists/P200000006165> (дата звернення: 18.12.2025).
3. Devore J. L. *Probability and Statistics for Engineering and the Sciences*. – 9th ed. – Boston: Cengage Learning, 2016. URL: <https://www.cengage.com/c/probability-and-statistics-for-engineering-and-the-sciences-9e-devore/> (дата звернення: 18.12.2025).
4. McKinney W. *Python for Data Analysis: Data Wrangling with pandas, NumPy, and IPython*. – 2nd ed. – Sebastopol: O'Reilly Media, 2018. URL: <https://www.oreilly.com/library/view/python-for-data/9781491957653/> (дата звернення: 19.12.2025).
5. Han J., Kamber M., Pei J. *Data Mining: Concepts and Techniques*. – 3rd ed. – Waltham: Morgan Kaufmann, 2012. URL: <https://www.sciencedirect.com/book/9780123814791/data-mining-concepts-and-techniques> (дата звернення: 19.12.2025).

Василик Дарія Миколаївна – студент групи 2ПІ-246, Вінницький національний технічний університет, м. Вінниця, e-mail: dasha.vasilik0708@gmail.com

Кирилячук Світлана Анатоліївна – к. пед. н., доцент кафедри вищої математики, Вінницький національний технічний університет, м. Вінниця, e-mail: kyrylashchuk@vntu.edu.ua

Vasylyk Daria M. – student of 2BC-24b group, Vinnytsia National Technical University, Vinnytsya, e-mail: dasha.vasilik0708@gmail.com

Kyrylashchuk Svitlana A. – Candidate of Pedagogical Sciences, Associate Professor, Department of Higher Mathematics, Vinnytsia National Technical University, e-mail: kyrylashchuk@vntu.edu.ua