

НОРМАЛІЗАЦІЯ СЕНСОРНИХ ДАНИХ ЗА ДОПОМОГОЮ МЕТОДУ Z-SCORE ТА ХОДЖЕСА-ЛЕМАНА

Вінницький національний технічний університет

Дана робота присвячена дослідженню та порівняльному аналізу методів Z-score та Ходжеса-Лемана для нормалізації сенсорних даних, отриманих з безпілотних літальних апаратів (БПЛА). Метою дослідження є оцінка ефективності кожного методу у стандартизації даних, виявленні аномалій та покращенні загальної якості сенсорних даних, що є критично важливим для забезпечення надійної роботи БПЛА. Проведено порівняльний аналіз, де оцінюються переваги та недоліки кожного методу в контексті обробки даних з різних типів сенсорів, встановлених на БПЛА. Особливу увагу приділено впливу викидів та шуму на ефективність нормалізації, а також здатності методів зберігати важливу інформацію у даних. Результати дослідження демонструють, що Z-score є ефективним для нормалізації даних з нормальним розподілом, тоді як метод Ходжеса-Лемана показує кращі результати при наявності викидів та ненормальному розподілі. Показано, що вибір оптимального методу нормалізації залежить від характеристик сенсорних даних та конкретного застосування БПЛА. Отримані результати можуть бути використані для розробки більш ефективних систем обробки сенсорних даних БПЛА, що дозволить покращити точність аналізу та виявлення потенційних проблем. Ключовим напрямком подальших досліджень є розробка адаптивних алгоритмів, що автоматично вибирають оптимальний метод нормалізації залежно від характеристик даних.

Ключові слова: нормалізація даних БПЛА, сенсорні дані, Z-критерій, оцінка Ходжеса-Лемана, стандартизація даних, непараметричні методи, виявлення викидів.

Вступ

У сфері безпілотних літальних апаратів (БПЛА) сенсорні дані відіграють ключову роль, забезпечуючи основу для прийняття рішень у різноманітних застосуваннях, від моніторингу довкілля до інспекції інфраструктури. Однак необроблені сенсорні дані часто характеризуються різними шкалами вимірювання, шумом та наявністю викидів, що може ускладнити їхній безпосередній аналіз та інтерпретацію. У цьому контексті нормалізація даних постає як важливий етап попередньої обробки, що дозволяє привести дані до єдиного масштабу, зменшити вплив екстремальних значень та підвищити якість подальшого аналізу.

Серед наявних методів нормалізації особливе місце займають Z-критерій та оцінка Ходжеса-Лемана. Z-критерій, або стандартний бал, є широко використовуваним статистичним інструментом для стандартизації даних шляхом вимірювання кількості стандартних відхилень, на які значення відхиляється від середнього. Цей метод особливо ефективний для виявлення аномальних значень та порівняння даних з різних розподілів. З іншого боку, оцінка Ходжеса-Лемана є непараметричним методом, який забезпечує стійку оцінку центральної тенденції набору даних. На відміну від параметричних методів, оцінка Ходжеса-Лемана не вимагає припущень щодо розподілу даних, що робить її цінним інструментом при аналізі сенсорних даних БПЛА, які можуть не завжди відповідати нормальному розподілу.

У цьому звіті буде проведено аналіз методів Z-критерію та Ходжеса-Лемана в контексті нормалізації сенсорних даних, отриманих з БПЛА. Розглянемо теоретичні основи цих методів, їхні статистичні властивості та потенційні застосування для покращення якості та інтерпретованості даних дистанційного зондування [1-2].

Результати дослідження

Існує ряд фундаментальних критеріїв, які використовуються для оцінки та порівняння статистичних методів. Ці критерії допомагають зрозуміти придатність кожного методу для конкретних типів даних та дослідницьких питань. Параметричні методи, такі як Z-score, зазвичай використовуються для неперервних даних і часто передбачають певний розподіл (наприклад, нормальний), тоді як непараметричні методи, такі як метод Ходжеса-Лемана, є більш гнучкими і можуть застосовуватися до різних типів даних без суворих припущень щодо розподілу. Вибір між параметричними та непараметричними методами залежить від характеру даних та обґрунтованості припущень щодо розподілу.

Параметричний характер Z-score дозволяє робити сильніші висновки, якщо припущення виконуються, тоді як метод Ходжеса-Лемана пропонує робастну альтернативу, коли ці припущення є сумнівними. Розуміння шкали та характеру даних (наприклад, неперервні, категоріальні, нормально розподілені) має вирішальне значення для вибору відповідного методу.

Z-score значною мірою покладається на припущення про нормальність, особливо для інтерпретації ймовірностей та виявлення викидів на основі стандартних відхилень. Метод Ходжеса-Лемана, будучи непараметричним, робить менше припущень щодо розподілу, хоча симетрія важлива для його інтерпретації як оцінки різниці медіан. Обґрунтованість методу Z-score тісно пов'язана з нормальністю даних. Порушення цього припущення може призвести до оманливих висновків. Метод Ходжеса-Лемана має більш гнучкі вимоги до розподілу. Статистичні тести будуються на певних умовах щодо даних. Якщо ці умови не виконуються, результати тесту можуть бути ненадійними. Для Z-score припущення про нормально розподілену сукупність є фундаментальним для багатьох його застосувань. Метод Ходжеса-Лемана, з іншого боку, використовує ранги або парні порівняння, які менш чутливі до конкретної форми розподілу.

Метод Ходжеса-Лемана відомий своєю робастністю до викидів, тобто він менш чутливий до екстремальних значень порівняно з Z-score, який використовує середнє значення та стандартне відхилення, може бути значно вплинутий викидами. У наборах даних, схильних до викидів або екстремальних значень, метод Ходжеса-Лемана пропонує більш стабільну міру центральної тенденції або зсуву положення, ніж Z-score. Викиди можуть спотворювати середнє значення та стандартне відхилення, ключові компоненти Z-score. Це спотворення може призвести до неточних оцінок того, наскільки інші точки даних віддалені від центру розподілу. Метод Ходжеса-Лемана, зосереджуючись на медіанах та рангах, за своєю суттю зменшує вплив екстремальних значень, забезпечуючи більш надійну оцінку за їхньої наявності.

Ефективність стосується того, наскільки добре метод використовує дані для оцінки параметра. Для симетричних розподілів метод Ходжеса-Лемана може бути досить ефективним, іноді навіть ефективнішим за вибіркиму медіану. Ефективність Z-score є оптимальною, коли дані дійсно нормально розподілені. Хоча Z-score є високоефективним за нормальності, метод Ходжеса-Лемана забезпечує хорошу ефективність для ширшого діапазону розподілів, особливо симетричних. Статистична ефективність є мірою того, скільки інформації про параметр сукупності отримує оцінювач з даної вибірки. Для даних, які ідеально відповідають припущенням параметричного тесту, цей тест часто є найефективнішим. Однак, якщо припущення не повністю виконуються, робастний непараметричний оцінювач, такий як метод Ходжеса-Лемана, може запропонувати хороший баланс надійності та ефективності [3-4].

Z-score легко інтерпретувати як кількість стандартних відхилень від середнього значення. Метод Ходжеса-Лемана оцінює зсув положення або псевдомедіану, що може вимагати дещо іншої інтерпретації залежно від контексту (одна вибірка проти двох вибірок). Пряма інтерпретація Z-score як стандартних відхилень від середнього значення робить його легко зрозумілим. Інтерпретація методу Ходжеса-Лемана як медіани різниць або псевдомедіани може вимагати більше контексту для тих, хто не знайомий з непараметричною статистикою.

Ключовим аспектом будь-якого статистичного методу є здатність зрозуміти та передавати його результати. Z-score має дуже інтуїтивно зрозумілу інтерпретацію, пов'язану з розкидом даних навколо середнього значення. Результат методу Ходжеса-Лемана, будучи статистично значущим, може вимагати глибшого розуміння непараметричних концепцій для повного усвідомлення його наслідків [5].

Складність обчислення Z-score є досить простим, зазвичай $O(n)$. Метод Ходжеса-Лемана, особливо в його базовій формулі, може бути більш обчислювально-інтенсивним, потенційно $O(n^2 \log n)$ або $O(mn \log(mn))$, хоча існують більш ефективні алгоритми. Для дуже великих наборів даних нижча обчислювальна складність методу Z-score може бути перевагою.

Середнє значення та стандартне відхилення, основа Z-score, не є робастними статистиками, що робить метод виявлення викидів Z-score чутливим до самих викидів, які він має на меті ідентифікувати. Викиди за визначенням є значеннями, далекими від центральної тенденції. Z-score використовує середнє значення як міру центральної тенденції та стандартне відхилення як міру розкиду. Одне велике викидання може змістити середнє значення та збільшити стандартне відхилення, таким чином змінюючи Z-score всіх інших точок у наборі даних. Метод Ходжеса-Лемана: Робастний до викидів. Будучи заснованим на медіані парних середніх або різниць, він не зазнає значного впливу екстремальних значень. Сама медіана є робастною мірою. Оцінка Ходжеса-Лемана має відносно високу точку руйнування, що вказує на його стабільність навіть за значної частки викидів у даних. Обчислення оцінки Ходжеса-Лемана на основі медіани забезпечує природну стійкість до впливу екстремальних значень, що робить його більш надійною мірою положення або зсуву за наявності викидів. Медіана визначається як середнє значення у відсортованому наборі даних. Її положення не залежить від фактичних значень точок даних за межами медіани. Оскільки оцінка Ходжеса-Лемана виводиться з медіани парних значень, екстремальні викиди, впливаючи на діапазон i , можливо, на деякі парні значення, не мають істотного впливу на медіану цього набору, таким чином забезпечуючи робастність оцінки. [6]

Основна корисність Z-score полягає в його здатності стандартизувати дані та кількісно визначати відхилення від середнього значення, що робить його універсальним інструментом для виявлення незвичайних спостережень та порівняння даних за загальною шкалою в різних галузях. Застосування методу Z-score зумовлене необхідністю розуміння того, де знаходиться точка даних відносно середнього значення набору даних. Це має вирішальне значення в таких сценаріях, як виявлення дефектних продуктів у виробництві (контроль якості), виявлення шахрайських транзакцій у фінансах (виявлення викидів) або підготовка даних для моделювання шляхом забезпечення того, щоб усі ознаки були в порівнянному масштабі (стандартизація). Метод Ходжеса-Лемана: Зазвичай застосовується в клінічних дослідженнях для оцінки величини ефектів лікування (наприклад, медіанна різниця в результатах), особливо коли дані не мають нормального розподілу або містять викиди [7]. Він також використовується в аналізі виживаності для оцінки відмінностей у даних про час до події та у фармацевтичних дослідженнях для непараметричної оцінки зсувів положення. Його робастність робить його цінним у будь-якій галузі, де дані можуть відхилятися від ідеальних параметричних припущень.

Таблиця 1

Порівняльний аналіз Z-score та Ходжеса-Лемана

Характеристика	Z-score	Ходжеса-Лемана
Тип даних	Неперервні (інтервальні/відносні), ідеально нормально розподілені	Неперервні (інтервальні/відносні), без суворих припущень щодо розподілу
Параметричний/Непараметричний	Параметричний	Непараметричний
Чутливість до викидів	Висока чутливість	Робастний

Припущення щодо розподілу	Передбачає нормальність (особливо для ймовірнісної інтерпретації)	Менше припущень; симетрія відносно медіани важлива для деяких інтерпретацій
Оцінює	Стандартизоване значення відносно середнього та СВ; використовується в Z-тесті для середніх	Параметр положення (псевдомедіана або медіанна різниця/зсув)
Обчислювальна складність	$O(n)$	Близько $O(n \log n)$ з ефективними алгоритмами
Застосування	Виявлення викидів, стандартизація, перевірка гіпотез (Z-тест), композитні оцінки, фінанси, медичні оцінки, контроль процесів	Ефекти лікування в клінічних дослідженнях, не менша ефективність, аналіз виживаності, фармацевтична статистика, робастна оцінка положення

Висновки

Метод Z-score є потужним інструментом, коли дані є або вважаються нормально розподіленими, пропонуючи просту інтерпретацію для стандартизації та виявлення викидів на основі стандартних відхилень від середнього значення. Його обчислювальна ефективність також є значною перевагою для великих наборів даних. Однак його чутливість до викидів та залежність від припущення про нормальність можуть бути обмеженнями при роботі з реальними даними, які часто відхиляються від цих ідеальних умов. З іншого боку, оцінка Ходжеса-Лемана забезпечує робастну непараметричну альтернативу, яка менш чутлива до викидів і не вимагає сильних припущень щодо розподілу. Це робить його особливо цінним у таких галузях, як клінічні дослідження та аналіз виживаності, де дані можуть бути скошеними або містити екстремальні значення. Хоча він, як правило, є більш обчислювально-інтенсивним, ніж Z-score, ефективні алгоритми роблять його застосовним у багатьох практичних сценаріях.

Слід зазначити, що існують потенційні модифікації або пов'язані методи, які можуть усунути деякі обмеження кожного підходу. Наприклад, при роботі з ненормальними даними та виявленні викидів модифікований Z-score з використанням медіанного абсолютного відхилення (MAD) замість стандартного відхилення може бути більш робастним. Подібним чином, для дуже великих наборів даних, де обчислювальні витрати методу Ходжеса-Лемана можуть бути проблемою, дослідники можуть розглянути наближення або альтернативні робастні оцінювачі.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Rebecca Bevans, "Choosing the Right Statistical Test | Types & Examples", January 28, 2020, <https://www.scribbr.com/statistics/statistical-tests/>
2. P. Mishra, C. Pandey, U. Singh, A. Keshri, M. Sabaretnam, "Selection of Appropriate Statistical Methods for Data Analysis", https://doi.org/10.4103/aca.ACA_248_18
3. How to Choose the Right Statistical Test | Types and Examples - Enago, <https://www.enago.com/academy/right-statistical-test/>
4. I. Alam, "Detecting Outliers Using Z-score — part -2", <https://medium.com/@irshadalamtech/detecting-outliers-using-z-score-part-2-85297f54ea84>
5. Z-Scores vs. Box Plots: The Quirky Quest to Unearth Outliers in Statistical Wonderland | by Patwariraghattam | DevOps.dev, <https://blog.devops.dev/z-scores-vs-box-plots-the-quirky-quest-to-unearth-outliers-in-statistical-wonderland-cb2f04>

[89984e](#)

6. Gerd K. Rosenkranz, "A note on the Hodges-Lehmann estimator", 15 June 2010, <https://doi.org/10.1002/pst.387>
7. B. M. Brown, D. G. Kildea, "Reduced U-Statistics and the Hodges-Lehmann Estimator" Ann. Statist. 6(4): 828-835 (July, 1978). DOI: 10.1214/aos/1176344256

Проценко Михайло Ігорович — аспірант кафедри АІТ, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: mishagg45@gmail.com

Маслій Роман Васильович – доцент кафедри АІТ, Вінницький національний технічний університет, Вінниця, e-mail: maslij.r.v@vntu.edu.ua

M. I. Protsenko
R. V. Maslii

NORMALIZATION OF SENSOR DATA USING THE Z-SCORE AND HODGES-LEHMAN METHODS

Vinnitsia National Technical University

This paper is devoted to the study and comparative analysis of the Z-score and Hodges-Lehman methods for normalizing sensor data obtained from unmanned aerial vehicles (UAVs). The study aims to assess each method's effectiveness in standardizing data, detecting anomalies, and improving the quality of sensor data, which is critical for ensuring reliable UAV operation. A comparative analysis is carried out to assess the advantages and disadvantages of each method in the context of processing data from different types of sensors installed on UAVs. Particular attention is paid to the impact of emissions and noise on normalization performance, as well as the ability of the methods to preserve important information in the data. The results of the study demonstrate that the Z-score is effective for normalizing data with a normal distribution, while the Hodges-Lehman method shows better results in the presence of outliers and non-normal distribution. It is shown that the choice of the optimal normalization method depends on the characteristics of the sensor data and the specific application of the UAV. The results obtained can be used to develop more efficient UAV sensor data processing systems, which will improve the accuracy of analysis and identify potential problems. A key area for further research is the development of adaptive algorithms that automatically select the optimal normalization method depending on the characteristics of the data.

Keywords: UAV data normalization, sensor data, Z-score, Hodges-Lehman estimation, data standardization, non-parametric methods, emission detection.

Protsenko Mykhailo I. — Department of Intelligent Information Technologies and Automation, Vinnitsia National Technical University, Vinnitsia, e-mail: mishagg45@gmail.com

Maslii Roman V. – associate professor at the Department of AIIT, Vinnitsia National Technical University, Vinnitsia, email: maslij.r.v@vntu.edu.ua