Математика

УДК 519.24

**Рибалка Антон Миколайович**

*студент*

*Національного технічного університету України*

*«Київський політехнічний інститут імені Ігоря Сікорського»*

# МЕТОДИ ОБРОБКИ ЕМПІРИЧНИХ ДАНИХ НА ОСНОВІ ЧИСЕЛЬНОГО ІМОВІРНІСНОГО АНАЛІЗУ

***Анотація.*** *В даній роботі було проаналізовано наукові публікації стосовно методів обробки емпіричних даних. Базуючись на розглянутих матеріалах було проведено порівняння найвідоміших способів обробки даних та представлено модернізацію одного з методів чисельного імовірнісного аналізу, а саме методу усереднення зміщених гістограм, який допомагає покращити обробку емпіричних даних в порівнянні з іншими способами.*

***Ключові слова:*** *емпіричні дані, чисельний імовірнісний аналіз, густина ймовірності, відновлення функції густини ймовірності, гістограма.*

***Summary.*** *In this paper, scientific publications on methods of empirical data processing were analyzed. Based on the considered materials, the comparison of the most known methods of data processing was carried out and the modernization of one of the methods of numerical probabilistic analysis was presented, namely the method of averaging shifted histograms, which helps to improve empirical data processing in comparison with other methods.*

***Key words:*** *empirical data, numerical probability analysis, probability density, restoration of probability density function, histogram.*

**Вступ.** Аналіз досліджуваної області показав, що тема обробки емпіричних даних з отриманням максимально надійних і достовірних результатів в даний час є досить актуальною і має застосування в найрізноманітніших областях. Так, наприклад, у сучасній практиці можна виділити такі сфери, де часто доводиться обробляти результати багатократних вимірювань величин, як: служби контролю якості підприємств; виробництво і експлуатація дорогих і високонадійних технічних виробів. Аналогічні приклади можна спостерігати в медицині, біології, гідроенергетиці, ракетобудуванні [1]. Також, як правило, різні методи обробки емпіричних даних часто використовуються не тільки в точних науках, а й в таких напрямках як психологія, соціологія, педагогіка, тощо.

У цих та інших областях доводиться обробляти великі об'єми вхідних даних, таких як сигнали декількох датчиків, що вимірюють одну і ту ж величину (температура, тиск, зміна стану об'єкта або навігаційні параметри і т. д), різні опитування, інтерв'ю. У процесі роботи з такого роду системами необхідно вирішувати велику кількість завдань, пов'язаних з управлінням надійністю, оцінкою опірності технічних пристроїв, прогнозуванням відмов технічних виробів та багато іншого.

Вирішення описаних вище задач характеризується високим рівнем невизначеності. Як об'єкт дослідження в роботі виступають емпіричні дані. Доведено, що найбільш повна інформація про об'єкт дослідження в умовах невизначеності на основі отриманих емпіричних даних виходить шляхом відновлення густини ймовірності невідомої випадкової величини [2]. Отже, в ролі предмета дослідження виступають методи відновлення густини ймовірності.

Головну мету даної роботи можна сформулювати як підвищення ефективності і якості оцінки стану систем в умовах невизначеності,

використовуючи обробку емпіричних даних на основі чисельного імовірнісного аналізу

**Постановка задачі.** Основою вибору оптимального рішення в умовах невизначеності є відновлення функції густини ймовірності. Оскільки дана задача актуальна для широкого спектру прикладних наук та галузей підприємства, сформувалося багато методів вирішення даної задачі. Проте вони всі мають недоліки, такі як неточність результатів та складність реалізації на практиці [3]. Щоб вирішити ці проблеми було запропоновано спосіб модернізації методу усереднення зміщених гістограм Девіда Скотта.

**Чисельний імовірнісний аналіз**

В ході вивчення різного роду систем доводиться стикатися з деякими труднощами, однією з яких є недостатня кількість інформації про об'єкт та його процеси. В результаті чого необхідно вирішувати завдання обробки інформацією з метою подальших досліджень в умовах невизначеності. На якість прийнятих рішень буде впливати отриманий рівень знань, який буде визначатися різними підходами і методами. Крім того, особливе значення мають методи та способи обробки цієї інформації, що відрізняються між собою складністю реалізації. Тому виникає ще одне завдання розробки методу, який дозволить не тільки збільшити достовірність інформації, але і знизити рівень складності реалізації самого методу, щоб не тільки заощадити часові та трудові ресурси, а й виключити можливість помилки за рахунок реалізації складного алгоритму обробки. Таке рішення може бути знайдено завдяки чисельному імовірнісному аналізу.

Чисельний імовірнісний аналіз – розділ обчислювальної математики, що займається вирішенням завдань з випадковими вхідними даними. Вивчення об'єктів і систем з використанням чисельного імовірнісного аналізу – особливо важливий і корисний інструмент в умовах невизначеності і ризику. Предметом чисельного імовірнісного аналізу є

рішення різних завдань зі стохастичними невизначеностями в даних з використанням чисельних операцій над щільністю ймовірностей випадкових величин і функцій з випадковими аргументами. Для цього пропонуються різноманітні інструменти, які включають такі поняття, як гістограмна арифметика, ймовірнісні і гістограмі розширення, гістограми другого порядку.

Чисельний імовірнісний аналіз являє собою непараметричний підхід і може успішно застосовуватися для ймовірнісного опису систем в рамках інтелектуально-інтератктивного моделювання, тим самим підвищуючи якість дослідження системи. З ціллю зниження рівня невизначеності та отримання додаткових даних про розподіл параметрів, пропонується застосувати гістограмний підхід.

**Функція густини ймовірності та її відновлення**

Функція густини ймовірності – це густина з якою розподіляється значення випадкової величини в конкретній точці. Функція густини імовірності існує лише для абсолютно неперервних випадкових величин.

Випадкова величина — величина, можливими значеннями якої є результати випробувань чи спостережень явищ або процесів, що носять випадковий характер.

Випадкову величину $\varepsilon$ називають абсолютно неперервною, якщо її функція розподілу допускає представлення:

$$F_\varepsilon(x) = \int_X \hat{f}_\varepsilon(x)dx,$$

де $\hat{f}_\varepsilon(x)$ – невід'ємна інтегрована за Лебегом функція.

Функція $\hat{f}_\varepsilon(x)$ називається функцією густини імовірності випадкової величини $\varepsilon$ . Вона є досить важливою характеристикою випадкової величини, тому проблема оцінки даної функції досить важлива. Відновлення густини ймовірності дозволить отримати надійний опис системи з максимально достовірним результатом.

Задача відновлення густини ймовірності по виборці є основною проблемою математичної статистики [4]. В більшості випадків, точний вид закону розподілу генеральної сукупності даних невідомий, тому необхідно відновлювати густину ймовірності по даній виборці. Дана задача (відновлення густини ймовірності) має практичне значення в багатьох областях, таких як наука, статистика, медицина, біологія. Вона допомагає отримати максимально надійний опис параметрів системи.

Існує багато непараметричних методів і алгоритмів відновлення густин ймовірності. Всі вони, в тій чи іншій мірі, мають свої недоліки та переваги. Серед найпопулярніших методів можна виділити:

- метод гістограм
- полігон частот
- метод ядерних оцінок

Розглянемо більш детально дані методи відновлення густини ймовірності.

**Аналіз існуючих методів відновлення густини ймовірності**

Метод гістограм розглянуто у роботах [5; 6]. Гістограма – це геометричне зображення емпіричної функції густини деякої випадкової величини, що побудоване по виборці даних. Висота стовбця вказує на частоту появи значень у обраному діапазоні, а кількість стовбців – на кількість діапазонів. Гістограма – це один з найпримітивніших та найстаріших методів відновлення густини ймовірності. Особливо важлива перевага гістограм полягає в тому, що вони добре візуалізують дані, а також цей метод досить простий в реалізації. Проте гістограма дає лише приблизне представлення про функцію густини ймовірності. А також при малих розмірах вибірки присутня проблема розбиття множини значень випадкової величини на інтервали.

Полігон частот – також один із способів представлення густини ймовірності. Він використовується для представлення неперервного та

дискретного розподілу [7]. Якщо розподіл неперервний і графік розподілу описується плавною залежністю, то полігон частот є кращим методом ніж гістограма, проте в загальному випадку, він також дає лише приблизне представлення про функцію густини ймовірності.

Метод ядерних оцінок було розглянуто у роботі [8]. Він ґрунтується на методі ядерного згладжування. Існує різна кількість методів згладжування, наприклад, такі як сплайни, але в асимптотичному сенсі вони є еквівалентні до методу ядерного згладжування. Процедура оцінки густини ймовірності в одній точці виглядає наступним чином:

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{t - T_i}{h}\right),$$

де $T$ – вибірка, яка була отримана в результаті спостереження за об'єктом;

$K$ – статичне ядро – симетрична, але не обов'язково додатна функція з інтегралом рівним одиниці;

$h$ – діапазон, параметр згладжування, який впливає на точність оцінок;

$n$ – розмір вибірки.

В якості ядра використовується Гаусове ядро:

$$K(u) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{\frac{1}{2}u^2},$$

де $u$ – значення, яке приймає дана функція.

Таким чином, густина ймовірності в точці $X$ обчислюється як сума значень ядра для величин, які визначаються між значенням $X$ і значенням послідовності. При цьому, точки $X$ можуть не співпадати зі значення самої послідовності. Даний метод є одним із найточніших методів відновлення густини ймовірності, проте, на практиці його доволі важко реалізувати, в результаті це нерідко призводить до помилок при обробці даних. Як

бачимо, розглянуті вище методи є недосконалими, а, враховуючи, необхідність якомога точніших та надійних розрахунків у деяких галузях (наприклад, медицина), розвиток методів обробки емпіричних даних є доволі значущим.

**Модернізований метод усереднення зміщених гістограм**

За основу було взято алгоритм запропонований Девідом Скоттом, а саме метод усереднення зміщених гістограм [9]. Він дозволяє враховувати всі наявні пустоти між стовпцями частотного полігону, зберігаючи при цьому обчислювальні переваги оцінки щільності розподілу. Щоб вирішити проблему згладжування, Скотт запропонував метод усереднення декількох зміщених гістограм. В ході експерименту параметри згладжування залишалися незмінними, але при цьому змінювалися початкові точки побудови стовпців. За словами вченого, даний метод досить добре підходить для оцінки щільності розподілу, в порівнянні з іншими методами. Розглянемо більш детально метод усереднення зміщених гістограм.

Даний метод полягає у тому, що початкова точка $x_0$ обирається m разів на певному відрізку. На кожній ітерації відбувається зміщення гістограми відносно осі x на $\frac{h}{m}$ одиниць, де $h$ – це ширина кожного стовпця. Тобто на першій ітерації початкова точка $x_{0_1} = x_0$, на другій

$x_{0_1} = x_0 + \frac{h}{m}$, маємо:

$x_{0_m} = x_0 + \frac{(m-1)*h}{m}$.

В такому випадку вихідна усереднена густина ймовірності визначається як середнє арифметичне значення густин гістограм на кожній ітерації:

$$\hat{f}_{\text{сер}}(x) = \frac{1}{m}\sum_{i=1}^{m}\hat{f}_i(x),$$

де – $\hat{f}_{\text{сер}}(x)$ усереднена густина ймовірності;

$\hat{f}_i(x)$ густина гістограми на кожній ітерації.

При цьому, на кожному із інтервалів $\frac{h}{m}$ функція є кусково-лінійною. Основою для використання даного методу для оцінки густини ймовірності є інтегральна збіжність частоти $(\frac{m}{N})$ до густини ймовірності:

$$\frac{m_i}{N} \to \int_{x_{i-1}}^{x_i} \hat{f}(x)dx,$$

де $N$ – розмір вибірки;

$\hat{f}(x)$ – густина ймовірності.

Даний метод має схожі показники результативності, в порівнянні з методом ядерних оцінок, проте набагато простіший в реалізації.

На основі публікацій Скотта запропонуємо модернізований метод усереднення зміщених гістограм. Вдосконалення полягає в тому, що оновленому алгоритмі будуть використовуватися тільки середні значення стовпців діаграми та потім буде проведено згладжування. На практиці це допоможе пришвидшити роботу метода та в певній мірі позбутися від надлишкових даних (випадкового зростання та (або) спадання даних при спостереженнях).

**Висновки.** В даній роботі було проаналізовано наукові роботи по методам обробки емпіричних даних. Було розглянуто, описано та порівняно популярні методи обробки емпіричних даних на основі чисельного імовірнісного аналізу, а саме методи відновлення густини імовірності, такі як метод гістограм, полігон частот та метод ядерних оцінок.

В результаті було модернізовано метод усереднення зміщених гістограм, порівняно його з іншими аналогічними методами. Даний результат дозволить підвищити ефективність і якість оцінки стану технічних систем, виробів і досліджень в умовах невизначеності.

**Література**


1. Вовк А. А., Основы общей теории статистики. Москва, 2006. С. 240.
2. Колмогоров А. Н., Основные понятия теории вероятностей. Москва, 1974. С. 119.
3. Корчикова Д. И., Арифметики и численный вероятностный анализнеопределенных данных. 2015.
4. Вапник В. Н., Восстановление зависимостей по эмперическим данным. 2007. С. 267.
5. Кропотов Ю. А., Методы оценивания моделей плотности вероятностей акустических сигналов в телекоммуникациях аудиообмена. 2017. С. 31-34. URL: http://sccs.intelgr.com/archive/2017-01/03-Kropotov.pdf
6. Глаголев М.В., Сабреков А.Ф., О восстановлении плотности вероятности методом гистограмм в почвоведении и экологии. 2010. С. 55-56. URL: https://www.ugrasu.ru/education/institutions/rec-environmental-dynamics-and-global-climate-change-the-unesco-chair/UNESCO_journal/docs/0/EDCC_0_55-83_Glagolev_Sabrekov.pdf
7. Блатов И.А., Старожилова О.В, Теория вероятностей и математическая статистика. Самара, 2010. С. 168.
8. URL: http://window.edu.ru/resource/896/74896/files/uch-pos-tv.pdf
9. Добронец Б. С. Численный вероятностный анализ неопределенных данных: монография / Б. С. Добронец, О. А. Попова. Красноярск: Сиб. фед. ун-т., 2014. С. 168.
10. Scott R. W., Multivariate density estimation: theory. practice. And visualization. John Wiley & Sons. 2015. С. 381.