ISSN 2219-9365

https://doi.org/10.31891/2219-9365-2023-76-8
УДК 004.94

ОДЕГОВ Микола
Державний університет інтелектуальних технологій і зв'язку
https://orcid.org/0000-0001-5526-2487
onick_64@ukr.net
БАБІЧ Юрій
Державний університет інтелектуальних технологій і зв'язку
https://orcid.org/0000-0002-7888-7591
babich159@gmail.com
БАГАЧУК Денис
Державний університет інтелектуальних технологій і зв'язку
https://orcid.org/0000-0001-8798-891X
bagachukdg@gmail.com
КОЧЕТКОВА Марина
Державний університет інтелектуальних технологій і зв'язку
jubdvg@gmail.com
ПЕТРОВИЧ Янна
Державний університет інтелектуальних технологій і зв'язку
yanna-petrovich@ukr.net

МЕТОДИКА ДВОКОМПОНЕНТНОГО ЕКСПРЕС-ТЕСТУВАННЯ НЕЗАЛЕЖНОСТІ ПОСЛІДОВНОСТЕЙ ПСЕВДОВИПАДКОВИХ ЧИСЕЛ

Для вирішення значної кількості задач методами імітаційного моделювання використовуються генератори випадкових та псевдовипадкових чисел. Повний аналіз статистичної незалежності пар ПВЧ треба виконувати за визначенням необхідних і достатніх умов: сумісні ймовірності дорівнюють добуткам маргінальних ймовірностей. Така схема призводить до необхідності вирішувати задачі у двомірному просторі, а порядок алгоритмів складає NxN, де N – довжина ПВЧ.

У даній роботі пропонується методика застосування двокомпонентного тесту, який включає тестування на відповідність функції розподілення сум ПВЧ теоретичній функції розподілення відповідних сум незалежних випадкових величин, а також тест некорельованості. Сумісне застосування цієї пари тестів дозволяє з достатньою практичною впевненістю відрізняти залежні та незалежні ПВЧ.

Ключові слова: випадкові числа, імітаційне моделювання, кореляція, статистична незалежність, генератори випадкових чисел, рівномірне розподілення

ODEGOV Nick, BABICH Yuri, BAHACHUK Denis,
KOCHETKOVA Marina, PETROVYCH Yanna
State University of Intellectual Technologies and Communication

TWO-COMPONENT INDEPENDENCE EXPRESS TESTING METHOD FOR SEQUENCES OF PSEUDO-RANDOM NUMBERS

Random and pseudo-random number generators are involved in solving a significant number of problems through simulation modeling. The sufficient randomness or unpredictability is essential for sequences of random numbers (RNSs). This paradigm of sufficient randomness is the basis for many test systems available now. We suggest that it is more appropriate to understand random variables in the tasks of random processes simulation as models of probability theory, where all properties of variables are determined by their distribution functions. A complete analysis of the RNSs pairs statistical independence must be carried out through defining the necessary and sufficient conditions: the joint probabilities are equal to the products of the corresponding marginal probabilities. This approach leads to the need of solving problems in a two-dimensional space with the NxN order of algorithms, where N is the length of the RNSs.

This work proposes a two-component test, that includes: a) testing the correspondence of the RNSs sums distribution function to the theoretical distribution function of independent random variables sums; b) a non-correlation test. The joint application of this tests pair allows to distinguish between dependent and independent RNSs with sufficient degree of certainty. In this case, the order of algorithms is only N, since the problems are solved in a one-dimensional space. Thus, the proposed approach allows to solve problems involving extremely large arrays of random numbers, i.e. the Big Data problems.

The proposed method of two-component testing is validated on samples of pseudo-random numbers generators of the NumPy library for the Python programming language. For the uniformly and normally distributed RNSs, the proposed method showed compliance with theoretical estimates of the wrong decisions probability, as well as a sufficiently high speed of data processing, which allows to recommend it be used in express tests of RNSs.

Keywords: random numbers, simulation modeling, correlation, statistical independence, random number generators, uniform distribution

Постановка проблеми у загальному вигляді та її зв'язок із важливими науковими чи практичними завданнями

Для вирішення значної кількості задач методами імітаційного моделювання використовуються генератори випадкових та псевдовипадкових чисел (ГВЧ) [1, 2]. Суттєвим для послідовностей випадкових чисел (ПВЧ) вважається «достатня випадковість», яка розуміється як «хаотичність», тобто «непрогнозованість». Саме виходячи з цієї парадигми «достатньої випадковості» розроблено системи тестів ГВЧ [3, 4]. Таке розуміння «випадковості» характерне та продуктивне для вирішення задач криптографії. У задачах моделювання випадкових процесів, на наш погляд, більш доцільно розуміння випадкових величин (ВВ) як моделей теорії ймовірностей, де ***всі властивості величин*** визначаються їх функціями розподілення ймовірностей (ФР) [5].

У задачах імітаційного моделювання застосовуються моделі типу «кольорового шуму» із заданими, наприклад, кореляційними функціями [6]. Втім, на наш погляд, із ПВЧ типу «білого шуму» можна відтворити і модель «кольорового шуму», а ось зворотна задача є досить складною. Тому саме тести незалежності ПВЧ мають суттєве значення для дослідження ГВЧ і не втрачають своєї актуальності, оскільки з розробкою нових мов програмування та нових прикладних бібліотек дана задача виникає знову і знову.

Найпростіший тест незалежності ПВЧ – перевірка на відсутність кореляції. Втім, це дуже слабкий тест, оскільки перевіряється лише лінійна незалежність. У більш складних тестах досліджуються моментні статистики більш високих порядків [7]. Абсолютна надійна перевірка незалежності ПВЧ може бути виконана лише у двовимірному просторі, де для ПВЧ довжиною N порядок алгоритмів складатиме N^2. Довжина неповторних ПВЧ, які генеруються програмними методами може складати астрономічні величини. При цьому досліджувати приходиться ПВЧ різної довжини, а це додатково збільшує кількість операцій. Тому при вирішенні задач тестування ГВЧ можуть виникати ситуації класу Big Data, коли надійні методи аналізу не дозволяють вирішити задачу у прийнятний час. У даному випадку іноді приходиться дещо нехтувати надійністю на користь швидкості [8].

Аналіз досліджень та публікацій

Метою даної роботи є обґрунтування методики експрес-тестування ГВЧ, заснованої на використанні подвійного критерію: критерію сум та критерію кореляцій ПВЧ.

Дана методика зводить задачу аналізу у просторі $N\times N$ до задачі аналізу в одномірному просторі. При цьому, зрозуміло, є поступки точності взамін на швидкість. Тому при тестуванні ГВЧ дану методику слід розуміти як сукупність експрес-тестів, які дозволяють вирішувати задачу на надвеликих масивах даних. У даній роботі, як приклад застосування методики, тестуються ГВЧ бібліотеки NumPy мови програмування Python.

Серед задач тестування ПВЧ умовно можна визначити два типи: базове тестування, та спеціальне тестування. Базове тестування в основному складають тести на відповідність ПВЧ тому чи іншому теоретичному закону розподілення, тобто задача вирішується в одновимірному просторі. Спеціальні тести [3, 4] вирішують більш складні задачі: пристосованості ГВЧ для вирішення практичних задач, як правило у багатовимірних просторах. У даній роботі ми будемо передбачати, що базове тестування ГВЧ вже успішно пройдене, а спеціальні тести будемо орієнтувати для задач типу моделювання «білого шуму» або «кольорового шуму» із заданими кореляційними функціями.

Формалізація методики

Відомо, що для незалежності ВВ X та Y необхідно і достатньо, щоб їх сумісне розподілення ймовірностей дорівнювало добутку маргінальних розподілень [9], що можна записати у такому загальному вигляді:

$$P(X,Y) = P(X)\cdot P(Y);\;\; F(x,y) = F_X(x)\cdot F_Y(y);\;\; f(x,y) = f_X(x)\cdot f_Y(y), \tag{1}$$

де P, F, f – відповідно розподілення ймовірностей, ФР ймовірностей та щільності розподілення, якщо останні визначені для даних ВВ. Для суми $Z = X + Y$ незалежних ВВ справедливі імплікації [9] для ФР:

$$F(x,y) = F_X(x)\cdot F_Y(y) \Longrightarrow F_Z(z) = \int_{-\infty}^{\infty} F_Y(z-x)dF_X = F_X * F_Y; \tag{2}$$

та для щільностей:

$$f(x,y) = f_X(x)\cdot f_Y(y) \Longrightarrow f_Z(z) = \int_{-\infty}^{\infty} f_Y(z-x)f_X(x)dx = f_X * f_Y; \tag{3}$$

або у загальному вигляді: $P(X,Y) = P(X)\cdot P(Y) \Longrightarrow P(X) * P(Y)$. Тобто, розподілення сум виражаються як ***згортки*** маргінальних розподілень. Важливо, що згортки, на відміну від двовимірних розподілень, є функціями лише ***однієї*** змінної – суми значень ВВ.

Вирішення задач перевірки незалежності методом порівняння теоретичних ФР сум ВВ з емпіричними ФР (ЕФР) сум пар ПВЧ, таким чином, дозволяє суттєво зменшити порядок алгоритмів аналізу:

вдвічі скорочується розмірність простору рішень. Втім, умови виду (2) або (3) є ***необхідними***, але ***не достатніми*** для вирішення задачі тестування ГВЧ на незалежність.

Тоді виникає така формальна задача: при яких ***додаткових*** умовах справедлива зворотна імплікація: $P(X) * P(Y) \Longrightarrow P(X) \cdot P(Y) = P(X,Y)$? Розглянемо цю задачу на прикладі найпростішого двовимірного розподілення.

Задача про дві розумні монетки

Вирішимо задачу шляхом «уявного моделювання». Отак, маємо дві монетки X та Y, які можуть випадати або гербом, або решіткою. Позначимо перший варіант як 0, а другий як 1. Нехай маргінальні розподілення для обох монеток рівномірні, тобто: $P_X(0) = P_X(1) = 0{,}5;\ P_Y(0) = P_Y(1) = 0{,}5$. Всього для кожної монетки існує лише $n = 2$ варіантів випадіння, а ймовірність кожного з них позначимо: $p = 1/n = 0{,}5$. Тоді, якщо ВВ X та Y незалежні, то двовимірне розподілення буде визначатися як матриця добутків:

$$P(k,m) = p^2, \qquad k,m = 0,1\,, \tag{4}$$

а згортка буде мати $2n-1$ можливих значень: 0, 1 та 2 з розподіленням ймовірностей:

$$P(0) = p^2 = 0{,}25;\ P(1) = 2p^2 = 0{,}5;\ P(2) = p^2 = 0{,}25. \tag{5}$$

Поряд з матрицею (4) для незалежних ВВ розглянемо матрицю з відхиленнями від ідеальної, можливо для залежних ВВ:

$$P(k,m) = p^2 + \Delta P(k,m), \qquad k,m = 0{,}1. \tag{6}$$

Нехай в результаті підсумовування ВВ з сумісним розподіленням (6) отримано те ж саме розподілення сум (5), що співпадає зі згорткою. Це може статися лише у випадку, коли справедлива система рівнянь:

$$\begin{cases} a)\ \Delta P(0{,}0) + \Delta P(1{,}0) + \Delta P(0{,}1) + \Delta P(1{,}1) = 0 \\ b)\ p^2 + \Delta P(0{,}0) = p^2 \\ c)\ 2p^2 + \Delta P(0{,}1) + \Delta P(1{,}0) = 2p^2 \\ d)\ p^2 + \Delta P(1{,}1) = p^2 \end{cases} \tag{7}$$

Перше рівняння (7.a) відповідає умові балансу міри і витікає з того, що сукупна ймовірність всіх можливих випадків має дорівнювати одиниці. За яких загальних умов для всіх відхилень внаслідок системи рівнянь (7) буде виконана умова незалежності (4): $\Delta P(k,m) = 0,\ k,m = 0{,}1$.?

Очевидні рішення дають рівняння (7.b) та (7.d): $\Delta P(0{,}0) = \Delta P(1{,}1) = 0$. Тоді рівняння (7.a) та (7.c) дають невизначеність, оскільки для виконання умови згортки (5) достатньо, щоб виконувалась умова:

$$\Delta P(0{,}1) = -\Delta P(1{,}0). \tag{8}$$

Залежних ВВ, для яких виконується умова (8) можна вигадати безліч. Наприклад, якщо $\Delta P(0{,}1) = -0{,}25$, а $\Delta P(0{,}1) = 0{,}25$, то умова згортки (5) все одно виконується, а умова незалежності (4) не виконується. Це лише зайвий раз підказує, що умова згортки є лише необхідною, а не достатньою умовою незалежності.

Накладемо додаткову умову ***симетрії*** матриці (6): $P(k,m) = P(m,k),\ k,m = 0{,}1$. Тоді система рівнянь

$$\begin{cases} p^2 + \Delta P(0{,}1) = p^2 + \Delta P(1{,}0) \\ \Delta P(0{,}1) = -\Delta P(1{,}0) \end{cases} \tag{9}$$

буде мати єдине тривіальне рішення: $\Delta P(0{,}1) = \Delta P(1{,}0) = 0$, тобто ***поєднання*** умов згортки та симетрії у даному випадку ***є достатньою*** умовою незалежності ВВ X та Y. При цьому жодна з цих умов окремо не є достатньою умовою незалежності ВВ. Втім, розглянемо більш загальний випадок.

Модель симетричних розподілень

Розглянемо систему двох ВВ X та Y з однаковими і симетричними відносно 0 маргінальними розподіленнями на сітці значень $-n, -n+1, \ldots, n$, тобто: $P_X(k) = P_X(-k), P_Y(k) = P_Y(-k), P_X(k), k = -n, \ldots, n$. Двовимірне розподілення у загальному випадку представимо у вигляді типу (6) – розподілення незалежних ВВ з відхиленнями:

$$P_{XY}(k,m) = P_X(k)P_Y(m) + \Delta P(k,m), \qquad k,m = -n, \ldots, n. \tag{10}$$

Для незалежних ВВ *X* та *Y* згортка у даному випадку буде мати одномірне розподілення:

$$P_Z(q) = \sum_{k+m=q} P_X(k)P_Y(m), \qquad q = 0,1,\dots,2n, \tag{11}$$

тоді у суми $Z = q$ будуть входити лише складові або парами $m = q - k$ та $k = q - m$, або ще з доданками $q = 2m = 2k$ в залежності від значення суми q.

Система, аналогічна системі (7) буде мати у даному узагальненому випадку $2n$ рівнянь при n^2 невизначених параметрах $\Delta P(k,m)$. При $n > 2$ така система буде недовизначеною, тобто може мати безліч тривіальних рішень, коли всі $\Delta P(k,m) = 0$. Але, якщо додати умови симетрії $P_{XY}(k,m) = P_{XY}(m,k)$, то по аналогії з рівняннями (9) отримаємо єдине тривіальне рішення: $\Delta P(k,m) = 0,\ k,m = -n,\dots,n$.

Таким чином, знову приходимо до висновку: умова відповідності розподілення сум ВВ *X* та *Y* згортці незалежних ВВ (***критерій згортки***) та умова симетрії матриці двомірного розподілення ймовірностей (***критерій симетрії***) разом ***достатні*** для незалежності цих ВВ.

Даний висновок розповсюджується і на випадок, якщо маргінальні математичні очікування M_X та M_Y не дорівнюють 0: на залежність або незалежність ВВ зміщення системи координат на якусь константу ніяк не впливає. Також не важко довести викладений вище принцип подвійного критерія: згортки + симетрії для безперервних ВВ. У даному випадку аналогічні результати можна отримати як наслідок з теореми Фубіні [9].

Отже, застосування критерія згортки зводить задачу аналізу у двовимірному просторі до задачі в одновимірному просторі. Якщо додатково довести, що двомірна матриця розподілення симетрична, то і ВВ *X* та *Y* будуть незалежними. Проблема, однак, полягає в тому, що встановити відповідність критерію симетрії не простіше, ніж напряму аналізувати незалежність ВВ у двовимірному просторі. Тому будемо обмежуватись лише встановленням факту симетрії маргінальних розподілень.

Кореляційний критерій симетрії

Розглянемо систему тих самих ВВ *X* та *Y* з тими ж самими розподіленнями, що і вище. Тоді, вочевидь, їх маргінальні математичні очікування: $M_X = 0,\ M_Y = 0$, а коваріація буде визначатись:

$$cov(X,Y) = \sum_k \sum_m k \cdot m \cdot P(k,m). \tag{12}$$

У сумах (12) для будь-якого значення m знайдеться пара значень k^* та $-k^*$, або значення $k^* = 0$. Якщо знову вимагати виконання умови симетрії, то $P(k^*,m) = P(-k^*,m)$, тоді

$$\forall m, \forall k:\ k \cdot m \cdot P(k,m) - k \cdot m \cdot P(-k,m) = 0 \implies cov(X,Y) = 0,$$

тобто некорельованість можна вважати ***ознакою***, точніше, ***необхідною*** умовою симетрії. Якщо до цього додати симетрію маргінальних розподілень відносно середнього, то це може бути суттєвим обґрунтуванням симетрії двомірних розподілень, хоча у строго математичному розумінні це не є фактом.

Критерії двокомпонентного експрес-тесту незалежності ПВЧ

Надалі обмежимось розглядом пар ПВЧ, теоретичними аналогами яких є безперервні ВВ з однаковими симетричними (відносно математичного очікування) маргінальними ФР. Тобто, передбачаємо, що базові тести ГВЧ пройшли успішно. Із попереднього витікає, що достатньо інформативним для визначення незалежності/залежності ПВЧ може бути система двох тестів:

– незалежності за критерієм сум;

– симетрії за критерієм некорельованості.

Якщо обидва тести вказують на незалежність, то цього може бути достатньо для вирішення значної кількості прикладних задач. Зрозуміло, що у строго математичному розумінні така система тестів не є достатньою для встановлення факту незалежності. Але у задачах класу Big Data така система може бути прийнятною як за ознакою швидкості, так і за ознакою практичної надійності прийнятного рішення. Висока швидкість алгоритмів у даному випадку витікає з того, що задача аналізу двовимірних розподілень зводиться до задачі аналізу функцій однієї змінної z – суми значень пари ПВЧ *X* та *Y*.

Для визначення незалежності за критерієм сум у нашому випадку доцільно застосовувати методи непараметричної статистики: це дозволяє аналізувати розподілення незалежно від виду ФР. Серед таких методів можна застосовувати алгоритми, засновані на статистиці Колмогорова-Смирнова (СКС), статистиці ω^2 та ін. [5].

Майже всі ці статистики зводяться до функцій $d_n(\rho)$, які залежать від відстані $\rho(F, F_n)$ між теоретичною ФР $F(z)$ ВВ z та ЕФР $F_n(z)$. Якщо відома теоретична ФР значень $P(d_n(\rho) \le d) = K(d)$, то для будь-якого значення ймовірності p_0 (квантилі по рівню p_0) можна визначити критичне значення $d(p_0) = K^{-1}(p_0)$. Тоді випадки, коли $d_n < d(p_0)$ трактуються як відповідність т. з. «нульовій» гіпотезі ($H_0: F_n \equiv F$), а протилежні випадки – як відповідність альтернативній гіпотезі ($H_1: F_n \neq F$). При цьому допускається статистична похибка на рівні «довірчі» ймовірності p_0. Дані положення є загальновідомими, втім трактовка гіпотез у нашому випадку дещо інша: гіпотеза H_0 буде відповідати ситуації, коли виконана ***необхідна*** умова незалежності ПВЧ *X* та *Y*, а гіпотеза H_1 – протилежній ситуації.

Певним недоліком непараметричних методів є складність аналітичних формул, що виражають залежність квантилів статистик від об'єму вибірок n. Тобто, якщо дослідник змінює критеріальні значення у складних циклах, то розрахунок відповідних довірчих ймовірностей може займати значний час. Так, СКС $K_n(d)$ заснована на метриці Чебишева $\rho_{Ch}(n) = \sup_z |F(z) - F_n(z)|$, має загальний вигляд $d_n(\rho) = \sqrt{n}\rho_{Ch}(n)$ і дуже складну ФР:

$$P(d_n \le d) = K(d) = \sum_{j=-\infty}^{\infty} (-1)^j \exp(-2j^2 d^2). \quad (13)$$

Втім, при близьких до 1 значеннях «довірчої» ймовірності p_0 та протилежною ймовірністю $p_1 = 1 - p_0$ (близькою до 0) зворотна до ФР (13) – квантильна функція апроксимується залежністю:

$$K^1(d) = d(p_0) \approx \sqrt{-0{,}5 \log\big(0{,}5(1 - p_{0)}\big)}\,. \quad (14)$$

Завдяки простоті апроксимації (14) далі будемо використовувати саме СКС, а також визначимо критичні значення метрики Чебишева по рівню квантиля p_0, виходячи із залежності (14):

$$\rho_{Ch}(p_0, n) \approx \sqrt{-0{,}5 \log(0{,}5(1 - p_0)/n}\,. \quad (15)$$

Обчислювальні схеми у даній роботі включають аналіз пар ПВЧ довжиною 100, 1000 та 10000 елементів. Для цих випадків у табл. 1 наводяться критичні значення $\rho_{Ch}(p_0, n)$ для квантилів p_0 порядку 0,9 (90%) та 0,99 (99%). При цьому протилежна ймовірність p_1 теоретично буде дорівнювати 0,1 та 0,01 відповідно.

Для прийняття рішень за критерієм некорельованості будемо використовувати відому апроксимацію ФР емпіричного коефіцієнта кореляції r_n: при великих значеннях *n* вона наближається до ФР нормального закону [10] з параметрами: $Mr_n = r$; $\sigma r_n = \sigma_r = (1 - r^2)/\sqrt{n}$, де r – теоретичне значення коефіцієнта кореляції. У нашому випадку аналізуються ПВЧ, для яких $r = 0$. Значення квантилів стандартного нормального розподілення по рівням 0,9 та 0,99 дорівнюють відповідно 1,262 та 2,326. Відповідні критичні значення $|r_n|(p_0, n)$ також наводяться у табл. 1.

Таблиця 1

Критичні значення показника сум (метрики Чебишева) та коефіцієнта кореляції

Довжина ПВЧ *n*	По рівню $p_0 = 0{,}9$ $(p_1 = 0{,}1)$		По рівню $p_0 = 0{,}99$ $(p_1 = 0{,}01)$	
	$\rho_{Ch}(n)$	$\lvert r_n \rvert$	$\rho_{Ch}(n)$	$\lvert r_n \rvert$
100	0,1224	0,1262	0,1628	0,2326
1000	0,0387	0,0399	0,0515	0,0736
10000	0,0122	0,0126	0,0163	0,0233

Тестування ГВЧ бібліотеки NumPy мови Python

План обчислювальних експериментів

Загальний алгоритм тестування зводиться до виконання циклу з *M* тестів. Досліджується незалежність/залежність пар ПВЧ *X* та *Y* довжиною *N* випадкових чисел. У кожному парному тесті визначаються:

– мінімальні (min), середні (mean) та максимальні (max) по *M* значення статистик ρ_{Ch} та $|r_n|$;

– відповідні значення частот $p^*(0{,}1)$ та $p^*(0{,}01)$ перевищення статистиками ρ_{Ch} та $|r_n|$ відповідних теоретичних значень згідно табл. 1;

– встановлюється час виконання всіх *M* тестів;

– моделі теоретичної ФР та всіх *M* ЕФР відображаються графічно для виконання додаткового візуального аналізу.

Тестування незалежності рівномірно розподілених ПВЧ (РР ПВЧ)

Генерування РР ПВЧ $U[0,1]$ у діапазоні [0,1] має фундаментальне значення, оскільки значна кількість ГВЧ базується на методі «зворотних функцій» [1]. При цьому ПВЧ з будь-якою заданою ФР $G(x)$ генеруються за правилом: $x = G^{-1}(U[0,1])$. Не важко довести математичне положення: нехай є дві ***незалежні*** РР ПВЧ $U[0,1]$ X та Y. Тоді і ПВЧ X^* та Y^* , отримані методом «зворотних функцій» також ***незалежні***. Таким чином, достатньо перевірити незалежність пар РР ПВЧ, щоб використовувати їх для генерування незалежних ПВЧ з будь-якою іншою теоретичною ФР.

Для тестування обрано ГВЧ, який реалізується функцією numpy.random.rand(N), де N – довжина РР ПВЧ. У цьому ГВЧ реалізовано рекурентний алгоритм, який ініціалізується таймером комп'ютера. Такий спосіб має ту перевагу, що додає «хаосу» у ПВЧ.

Приклади ЕФР РР ПВЧ для різних значень N наведені на рис. 1.

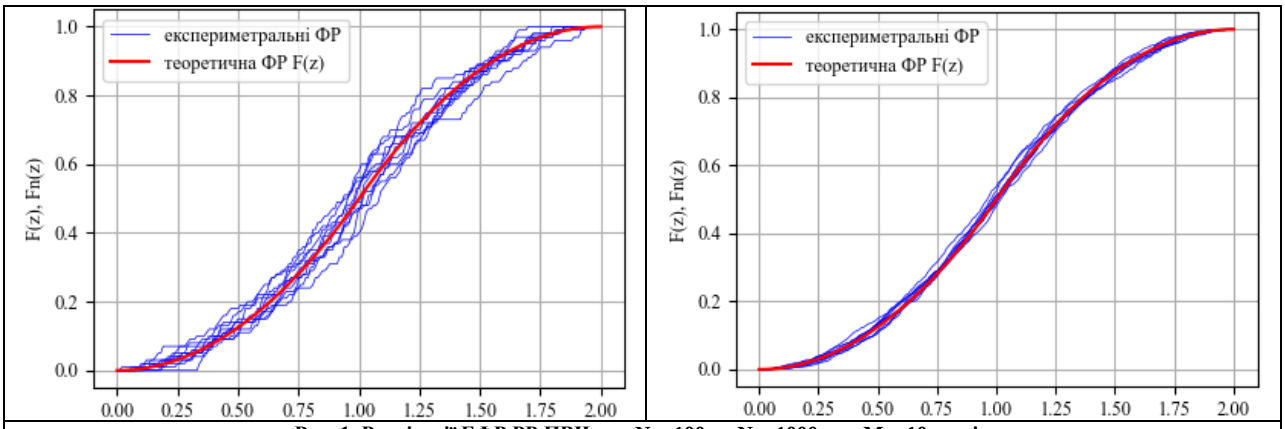

Рис. 1. Реалізації ЕФР РР ПВЧ для N = 100 та N = 1000 для M = 10 тестів

У першій серії тестів досліджувались неспотворені РР ПВЧ за гіпотезою незалежності. Основні результати тестування наведені у табл. 2 та 3. У табл. 2 наведені мінімальні, середні та максимальні значення статистик: метрика Чебишева (показник за критерієм сум) та коефіцієнта кореляції (показник за критерієм некорельованості). Аналіз цієї таблиці показує, що середні значення завжди знаходяться у межах справедливості гіпотези про незалежність, втім максимальні значення іноді перевищують критичні значення по табл. 1.

Результати досліджень перевищення критичних значень показані у табл. 3. У цій таблиці значення p^* відповідають емпіричній частоті виходу значень за критичні межі у серіях з M тестів.

Таблиця 2

Мінімальні, середні та максимальні значення статистик для ПВЧ довжини N у M тестах

N	M	Показник за метрикою Чебишева			Коефіцієнт кореляції		
		$\min \rho_{Ch}$	$\text{mean}\, \rho_{Ch}$	$\max \rho_{Ch}$	$\min\|r_n\|$	$\text{mean}\, r_n$	$\max\|r_n\|$
100	10000	0,0300	0,0820	0,2188	0,0000	0,0002	0,3811
1000	1000	0,0102	0,0259	0,0738	0,0000	-0,0007	0,1051
10000	100	0,0038	0,0084	0,0171	0,0000	0,0013	0,0272

Таблиця 3

Значення частот виходу за критичні межі для ПВЧ довжини N у M тестах

N	M	Показник за метрикою Чебишева		Модуль коефіцієнта кореляції	
		$p^*(0,1)$	$p^*(0,01)$	$p^*(0,1)$	$p^*(0,01)$
100	10000	0,0817	0,0077	0,2123	0,0212
1000	1000	0,0930	0,0060	0,2310	0,0170
10000	100	0,1100	0,0120	0,2100	0,0180

Аналіз даних у табл. 3 показує, що в усіх випадках за метрикою Чебишева значення $p^*(0,1)$ та $p^*(0,01)$ приблизно дорівнюють 0,1 та 0,01 відповідно, тобто добре узгоджуються з теоретичними критичними значеннями з табл. 1. Втім, ті ж самі значення для коефіцієнта кореляції приблизно вдвічі перевищують теоретичні. Причина проста: досліджувались не самі по собі коефіцієнти кореляції, а їх модулі. Тому, як не важко показати, частота виходу за критичні межі мала подвоїтись, що і сталось. З цих досліджень можна зробити ***суттєвий*** висновок: апроксимація (14) для квантилів статистики Колмогорова та нормальне наближення для коефіцієнта кореляції є достатньо точними для задачі, що вирішується.

В цілому, двокомпонентний тест за критеріями сум та некорельованості показує, що РР ВПЧ бібліотеки NumPy можна вважати достатньо незалежними для моделювання випадкових процесів типу білого шуму.

Перевірка роздільної здатності двокомпонентного тесту

У наступних серіях перевірялась роздільна здатність системи тестів на вочевидь залежних РР ПВЧ. У даних серіях методика виконувала протилежну задачу: встановити, що ПВЧ статистично залежні. Залежність у даних серіях моделюється керованою кореляцією. А саме, в парах ПВЧ X та Y якась частина α чисел ПВЧ Y замінюється числами з ПВЧ X. При цьому, зрозуміло: $0 \le \alpha \le 1$. Якщо визначити ціле число $m = [\alpha N]$, то ПВЧ Y^* буде мати структуру:

$$Y^*: y_n^* = x_n\,, n = \overline{1,m}\,;\;\; y_n^* = y_n\,, n = \overline{m+1,N}\,, \tag{16}$$

відповідно, послідовність сум Z буде мати структуру:

$$Z: z_n = 2x_n\,, n = \overline{1,m}\,;\;\; z_n = x_n + y_n, n = \overline{m+1,N}\,. \tag{17}$$

Порядок включення елементів ПВЧ X для заміщення елементів ПВЧ Y значення не має, але не важко довести, що у разі некорельованості ПВЧ X та Y коефіцієнт кореляції ПВЧ X та Y^* буде дорівнювати α. До речі: елементарні алгоритми програмування структур типу (16) дозволяють реалізувати моделі типу «кольорового шуму» із заданою кореляційною функцією.

Дані тестування при досить сильній кореляції ($\alpha = r = 0{,}5$) наведені у табл. 4. Як показує аналіз значень у цій таблиці, обидва тести (сум і некорельованості) при довжині ПВЧ N від 1000 впевнено вказують на наявність залежності: всюди значення p^* = 1 , тобто 100% ПВЧ не задовольняють критерію незалежності, що й треба у даному випадку. Але, для відносно малих вибірок порядку N = 100 тест сум дає спірні результати: лише приблизно 40% значень виходять за критичну межу. Висновок: парне застосування критерію сум та критерію некорельованості має добру розподільну здатність для вирішення задач дослідження незалежності потенційно корельованих ПВЧ.

Таблиця 4

Середні значення статистик та частот виходу за критичні межі для ПВЧ довжини N у M тестах (r = 0,5)

N	M	Показник за метрикою Чебишева			Коефіцієнт кореляції		
		$\text{mean}\,\rho_{Ch}$	$p^*(0{,}1)$	$p^*(0{,}01)$	$\text{mean}\,r_n$	$p^*(0{,}1)$	$p^*(0{,}01)$
100	10000	0,1174	0,3943	0,0709	0,4994	1,0000	0,9982
1000	1000	0,0763	1,0000	1,0000	0,4991	1,0000	1,0000
10000	100	0,0666	1,0000	1,0000	0,5016	1,0000	1,0000

Ще більш складний випадок, коли ПВЧ мають незначну кореляцію. Приклади ПВЧ з малою кореляцією порядку 0,2 показано на рис. 2. Як бачимо, при відносно невеликих довжинах ПВЧ ЕФР непогано «маскуються» під теоретичні ФР незалежних ВВ.

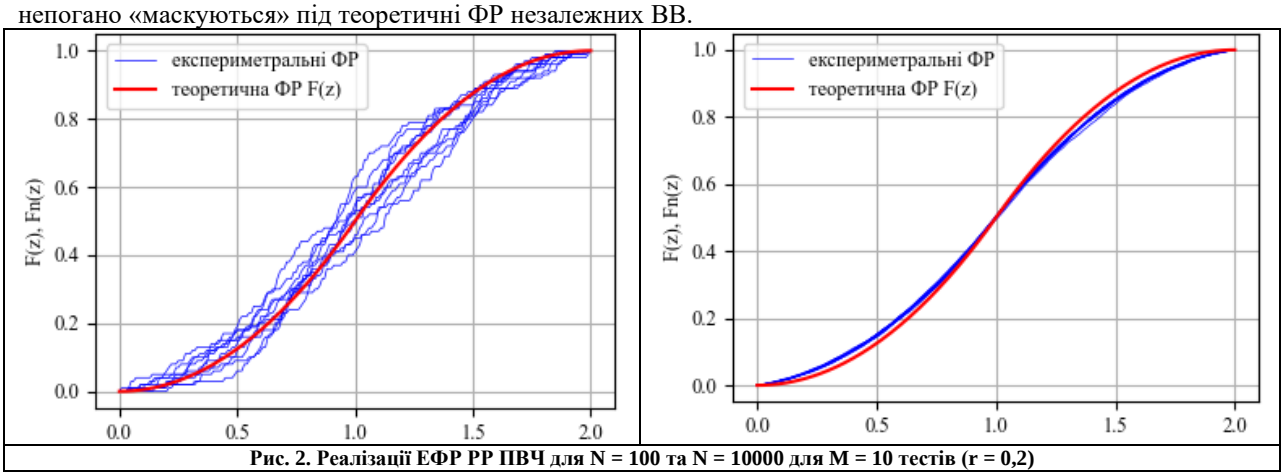

Рис. 2. Реалізації ЕФР РР ПВЧ для N = 100 та N = 10000 для M = 10 тестів (r = 0,2)

Дані тестування для цього випадку приведені у табл. 5. Як показує аналіз цієї таблиці, двокомпонентний тест у цьому разі також показує достатньо високу розподільну здатність: при N = 10000 обидва тести чітко вказують на наявність залежності, а при N = 1000 тест за критерієм сум відхиляє гіпотезу незалежності приблизно у 60% випадків.

Таблиця 5

Середні значення статистик та частот виходу за критичні межі для ПВЧ довжини *N* у *M* тестах (*r* = 0,2)

N	*M*	Показник за метрикою Чебишева			Коефіцієнт кореляції		
		$\text{mean}\,\rho_{Ch}$	$p^*(0{,}1)$	$p^*(0{,}01)$	$\text{mean}\,r_n$	$p^*(0{,}1)$	$p^*(0{,}01)$
100	10000	0.0895	0,1172	0,0111	0,1997	0,7739	0,3763
1000	1000	0.0414	0,5910	0,1340	0,2021	1,0000	1,0000
10000	100	0.0290	1,0000	1,0000	0,1999	1,0000	1,0000

Фактор часу рішення задач має суттєве значення для обробки вибірок значного обсягу. У табл. 6 показані результати виміру часу рішень при застосуванні запропонованого двокомпонентного тесту. Як видно з цієї таблиці, час прийняття рішень практично лінійно залежить від добутків $N{\times}M$, тобто від загальної кількості випадкових чисел, що аналізуються. В цьому й полягає основна перевага методики. Якщо напряму, за допомогою визначень (1) перевіряти незалежність пар ПВЧ, то час обробки даних зріс би астрономічно, орієнтовно в пропорції $(N{\times}M)^2$. У пропонованому алгоритмі рішення на масивах 2 000 000 приймались приблизно за 1,5 хвилини, що зовсім непогано для програм на Python, які виконуються під управлінням інтерпретатора. Аналогічні рішення за визначеннями (1) приймались би приблизно за 3 години. Коротко кажучи, коментарі зайві.

Таблиця 6

Час виконання тестів для ПВЧ довжини *N* у *M* тестах

N	*M*	*N*×*M*	*t*, с
100	100	10000	1,235
100	1000	100000	10,646
1000	100	100000	10.776
1000	1000	1000000	109,173
10000	100	1000000	107.322

Тестування незалежності нормально розподілених ПВЧ (НР ПВЧ)

Незалежні НР ПВЧ використовуються у багатьох задачах прикладного моделювання процесів типу «білого гаусового шуму». Якщо ГВЧ нормального розподілення заснований на методі зворотних функцій, то достатньо скористатись методикою для РР ПВЧ, які породжують ці НР ПВЧ. У даній роботі досліджуються НР ПВЧ, що генеруються за допомогою функції numpy.random.normal бібліотеки NumPy. Алгоритм цього ГВЧ (умовно) невідомий. Втім, теоретично маргінальні НР ПВЧ X та Y відповідають вимоги симетрії відносно математичних очікувань. Сума двох незалежних ВВ, які мають стандартне нормальне розподілення ($m_X = m_Y = 0$, $\sigma_X = \sigma_Y = 1$) буде мати таку ФР:

$$F(z) = \frac{1}{2\sqrt{\pi}} \int_{-\infty}^{z} \left[\exp\left(-\frac{z^2}{4} \right) \right] dz \qquad (18)$$

Приклади графіків ФР (18) та ЕФР для 10 тестів наведені на рис. 3.

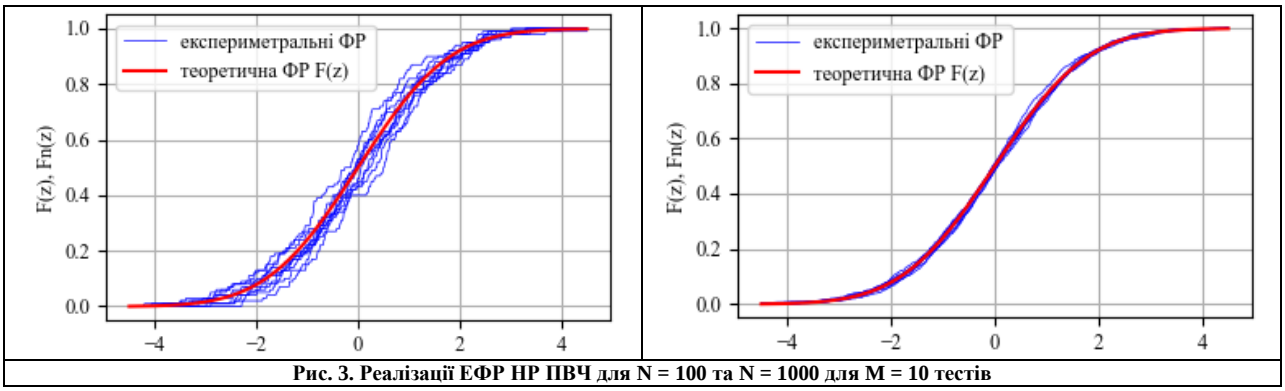

Рис. 3. Реалізації ЕФР НР ПВЧ для N = 100 та N = 1000 для M = 10 тестів

У табл. 7 дано основні результати тестування пар НР ПВЧ за допомогою запропонованого подвійного тесту.

Аналіз даних у табл. 7 в цілому співпадає з попередніми результатами для РР ПВЧ, саме:

– середні значення у *M* тестах чітко вказують на незалежність за обома критеріями;

– частота перевищення критичних значень по рівню 0,9 та 0,99 узгоджується з теоретичними значеннями;

– в цілому, НР ПВЧ бібліотеки NumPy можна вважати незалежними.

Таблиця 7

Середні значення статистик та частот виходу за критичні межі для НР ПВЧ довжини *N* у *M* тестах

N	*M*	Показник за метрикою Чебишева			Коефіцієнт кореляції		
		$\text{mean}\,\rho_{Ch}$	$p^*(0{,}1)$	$p^*(0{,}01)$	$\text{mean}\,r_n$	$p^*(0{,}1)$	$p^*(0{,}01)$
100	10000	0,0841	0,0871	0,0089	0,0000	0,2111	0,0188
1000	1000	0,0266	0,0760	0,0070	0,0012	0,2270	0,0230
10000	100	0,0091	0,1500	0,0200	0,0005	0,2000	0,0000

Висновки з даного дослідження і перспективи подальших розвідок у даному напрямі

1. Запропонований двокомпонентний тест (за критерієм сум та некорельованості) дозволяє звести задачу аналізу незалежності/залежності пар ПВЧ із двовимірного простору до одновимірного.

2. Дана методика найбільш ефективна для обробки надвеликих масивів ПВЧ (задачі класу Big Data), коли фактор часу прийняття рішень може по значущості навіть переважати фактор точності або надійності рішень.

3. ГВЧ бібліотеки NumPy Python, як для генерації РР ПВЧ, так і для генерації НР ПВЧ можуть бути запропоновані для імітаційного моделювання процесів типу «білого шуму».

4. Використані наближення критеріальних значень для обох тестів відповідають реальній статистиці і можуть бути запропоновані для експрес-тестування масивів ПВЧ надвеликої довжини.

5. Висока швидкість алгоритмів двокомпонентного тесту підтверджена експериментально: алгоритми мають порядок росту, що лише лінійно залежить від довжини ПВЧ.

Література

1. Буртняк І.В. Імітаційне моделювання / І.В. Буртняк. Прикарпатський національний університет ім. Василія Стефаника, 2019. – 97 с.

2. Герасимчук О.І. Генератори псевдовииадкових чисел, їх застосування, класифікація, основні методи побудови і оцінка якості / О.І.Герасимчук, В.М.Максимович // Науково - технічний журнал "Захист інформації", № 3, 2003. С. 29-36. DOI: https://doi.org/10.18372/2410-7840.5.4270.

3. Andrew Rukhin, JuanSoto, James Nechvatal, Miles Smid, ElaineBarker, Stefan Leigh, MarkLevenson, Mark Vangel, DavidBanks, Alan Heckert, Jame sDra and San Vo, " A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications: NIST Special Publication 800-22 Revision 1a", National Institute of Standards and Technology Gaithersburg, MD 20899-8930, Revised: April 2010. – 131 pp.

4. Roman Kochana, Lyudmila Kovalchuk, Oleksandr Korchenko and Nataliia Kuchynska, "Statistical Tests Independence Verification Methods", Procedia Computer Science, Volume 192, 2021, Pages 2678-2688, ISSN 1877-0509, DOI: https://doi.org/10.1016/j.procs.2021.09.038.

5. Карташов М.В. Імовірність, процеси, статистика / М.В. Карташов – Київ: ВПЦ Київський університет, 2008. – 494 с.

6. Juan Ignacio Deza, Hisham Ihshaish, "qNoise: A generator of non-Gaussian colored noise", SoftwareX, Volume 18, 2022, 101034, ISSN 2352-7110, DOI: https://doi.org/10.1016/j.softx.2022.101034.

7. Elena Almaraz Luengo, Marcos Brian Leiva Cerna, Luis Javier García Villalba and Julio Hernandez-Castro, "A new approach to analyze the independence of statistical tests of randomness", Applied Mathematics and Computation, Volume 426, 2022, 127116, ISSN 0096-3003, DOI: https://doi.org/10.1016/j.amc.2022.127116.

8. Одегов М.А. Обґрунтування швидких алгоритмів класифікації на множинах BIG DATA за критеріями надійності і продуктивності / М.А. Одегов, М.М. Гаджиєв, Л.М. Буката, Л.В. Глазунова, М.В. Кочеткова // Інфокомунікаційні та комп'ютерні технології. - №1, 2023. - С. 148 - 160. DOI: https://doi.org/10.36994/2788-5518-2023-01-05-16.

9. Ширяев А.Н. Вероятность / А.Н. Ширяев – М.: Наука, 1980. – 576 с.

10. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики – М.: Наука, 1983. - 416 с.

References

1. Burtnyak I.V. Imitacijne modelyuvannya / I.V. Burtnyak. Prikarpatskij nacionalnij universitet im. Vasiliya Stefanika, 2019. – 97 s.

2. Gerasimchuk O.I. Generatori psevdoviiadkovih chisel, yih zastosuvannya, klasifikaciya, osnovni metodi pobudovi i ocinka yakosti / O.I.Gerasimchuk, V.M.Maksimovich // Naukovo - tehnichnij zhurnal "Zahist informaciyi", № 3, 2003. S. 29-36. DOI: https://doi.org/10.18372/2410-7840.5.4270.

3. Andrew Rukhin, JuanSoto, James Nechvatal, Miles Smid, ElaineBarker, Stefan Leigh, MarkLevenson, Mark Vangel, DavidBanks, Alan Heckert, Jame sDra and San Vo, " A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications: NIST Special Publication 800-22 Revision 1a", National Institute of Standards and Technology Gaithersburg, MD 20899-8930, Revised: April 2010. – 131 rr.

4. Roman Kochana, Lyudmila Kovalchuk, Oleksandr Korchenko and Nataliia Kuchynska, "Statistical Tests Independence Verification Methods", Procedia Computer Science, Volume 192, 2021, Pages 2678-2688, ISSN 1877-0509, DOI: https://doi.org/10.1016/j.procs.2021.09.038.

5. Kartashov M.V. Imovirnist, procesi, statistika / M.V. Kartashov – Kiyiv: VPC Kiyivskij universitet, 2008. – 494 s.

6. Juan Ignacio Deza, Hisham Ihshaish, "qNoise: A generator of non-Gaussian colored noise", SoftwareX, Volume 18, 2022, 101034, ISSN 2352-7110, DOI: https://doi.org/10.1016/j.softx.2022.101034.

7. Elena Almaraz Luengo, Marcos Brian Leiva Cerna, Luis Javier Garcia Villalba and Julio Hernandez-Castro, "A new approach to analyze the independence of statistical tests of randomness", Applied Mathematics and Computation, Volume 426, 2022, 127116, ISSN 0096-3003, DOI: https://doi.org/10.1016/j.amc.2022.127116.

8. Odegov M.A. Obgruntuvannya shvidkih algoritmiv klasifikaciyi na mnozhinah BIG DATA za kriteriyami nadijnosti i produktivnosti / M.A. Odegov, M.M. Gadzhiyev, L.M. Bukata, L.V. Glazunova, M.V. Kochetkova // Infokomunikacijni ta komp'yuterni tehnologiyi. - №1, 2023. - S. 148 - 160. DOI: https://doi.org/ 10.36994/2788-5518-2023-01-05-16.

9. Shiryaev A.N. Veroyatnost / A.N. Shiryaev – M.: Nauka, 1980. – 576 s.

10. Bolshev L.N., Smirnov N.V. Tablicy matematicheskoj statistiki – M.: Nauka, 1983. - 416 s.