УДК 681.324.067 DOI: 10.31673/2412-4338.2022.034957

Поперешняк С.В.
Державний університет телекомунікацій, Київ

МЕТОДИКА СТАТИСТИЧНОГО АНАЛІЗУ ВИПАДКОВОСТІ ПОСЛІДОВНОСТЕЙ, ЩО ПОРОДЖУЮТЬСЯ ГЕНЕРАТОРАМИ ВИПАДКОВИХ ТА ПСЕВДОВИПАДКОВИХ ЧИСЕЛ

Анотація. *Проблеми Інтернету речей, зокрема необхідність захисту конфіденційних даних, призводять до використання методів шифрування і дешифрування, які, в свою чергу, покладаються на використання випадкових чисел. Використання ГПВЧ може допомогти забезпечити безпеку цього процесу, але важливо використовувати високоякісні ГПВЧ і ретельно їх тестувати, щоб уникнути вразливостей у безпеці. Статистичний аналіз використовується для оцінки випадковості згенерованих послідовностей шляхом тестування їх за допомогою статистичних тестів на випадковість. Ці тести оцінюють різні статистичні властивості, такі як розподіл частот згенерованих чисел, розподіл послідовних чисел і наявність закономірностей. Виконуючи статистичний аналіз, можна виявити будь-які недоліки або упередження в роботі генератора випадкових чисел і поліпшити його якість.*

В статті перевірено послідовність на випадковість за допомогою дво- та тривимірної статистики. Побудовано математичну модель, наведено теорему. Побудовано схему статистичного аналізу послідовностей в основі якої було використано методи програмної інженерії. Наведені приклади використання двійкової статистики для аналізу бітової послідовності.

Використання статистичного аналізу випадковості послідовностей, що генеруються генераторами випадкових (або псевдовипадкових) чисел, є популярною тенденцією, оскільки це має вирішальне значення для забезпечення надійності та безпеки різних комп'ютерних систем і додатків, включаючи Інтернет речей (IoT), криптографію, моделювання та наукові обчислення. Зі зростанням важливості кібербезпеки та конфіденційності даних попит на надійні та безпечні генератори випадкових чисел зростає, а використання статистичного аналізу для оцінки випадковості послідовностей, згенерованих генераторами випадкових (або псевдовипадкових) чисел, стає все більш критичним.

Ключові слова: *генератори випадкових та псевдовипадкових послідовностей, статистичний аналіз, бітова послідовність, Інтернет речей, статистичний аналіз.*

Popereshniak S.V.
State University of Telecommunications, Kyiv

METHOD OF STATISTICAL ANALYSIS OF RANDOM SEQUENCES GENERATED BY GENERATORS OF RANDOM AND PSEUDO-RANDOM NUMBER

Abstract. *The problems of the Internet of Things, in particular the need to protect confidential data, lead to the use of encryption and decryption methods, which, in turn, rely on the use of random numbers. The use of HVDC can help ensure the security of this process, but it is important to use high-quality HVDC and thoroughly test them to avoid security vulnerabilities. Statistical analysis is used to assess the randomness of the generated sequences by testing them using statistical tests of randomness. These tests evaluate various statistical properties, such as the frequency distribution of the generated numbers, the distribution of consecutive numbers, and the presence of patterns. By performing a statistical analysis, you can identify any flaws or biases in the operation of the random number generator and improve its quality.*

The article checks the sequence for randomness using two- and three-dimensional statistics. A mathematical model is built, a theorem is given. A scheme of statistical analysis of sequences was built, based on which software engineering methods were used. Examples of using binary statistics for bit sequence analysis are given.

The use of statistical analysis of the randomness of sequences generated by random (or pseudo-random) number generators is a popular trend, as it is critical to the reliability and security of various computing systems and applications, including the Internet of Things (IoT), cryptography, simulations, and scientific calculation. With the growing importance of cybersecurity and data privacy, the demand for reliable and secure random number generators is increasing, and the use of statistical analysis to assess the randomness of sequences generated by random (or pseudo-random) number generators is becoming increasingly critical.

Keywords: *generators of random and pseudo-random sequences, statistical analysis, bit sequence, Internet of things, statistical analysis.*

1. Постановка проблеми

Інтернет речей (IoT) має потенціал для революції в багатьох галузях і трансформації способу нашого життя та роботи. Генератори псевдовипадкових чисел (ГПВЧ) відіграють важливу роль у багатьох додатках IoT і забезпечують зручний та безпечний спосіб генерування випадкових значень. Статистичний аналіз випадковості послідовностей, згенерованих генераторами випадкових (або псевдовипадкових) чисел, використовується для оцінки якості та надійності цих генераторів. Випадковість послідовності є критично важливою властивістю генераторів випадкових чисел, оскільки вона безпосередньо впливає на їхню придатність для застосувань, де потрібні випадкові числа.

У багатьох випадках генератори випадкових чисел використовуються в таких чутливих додатках, як криптографія або наукове моделювання. Тому важливо переконатися, що згенеровані послідовності чисел є дійсно випадковими або, принаймні, мають достатній рівень статистичної випадковості, щоб зробити їх придатними для передбачуваного застосування.

Статистичний аналіз дозволяє кількісно оцінити випадковість згенерованих послідовностей і виявити будь-які закономірності або упередження, які можуть існувати. Піддаючи послідовності різним статистичним тестам, можна визначити ступінь їхньої випадковості та виявити будь-які недоліки в алгоритмі генератора.

Загалом, використання статистичного аналізу при оцінці генераторів випадкових (або псевдовипадкових) чисел гарантує, що згенеровані послідовності мають високу якість і підходять для передбачуваного застосування.

2. Аналіз останніх досліджень і публікацій.

Існує багато вчених, які зробили значний внесок в область статистичного аналізу випадковості послідовностей, що генеруються генераторами випадкових (або псевдовипадкових) чисел. Ось декілька прикладів:

Джордж Марсалья: Марсалья був математиком і статистиком, який зробив багато внесків у сферу генерації випадкових чисел. Він розробив кілька відомих ГПСЧ, включаючи генератори XORshift і MWC. Він також створив кілька статистичних тестів для оцінки якості генераторів випадкових чисел, включаючи тести Diehard і BigCrush.

Дональд Кнут: Кнут - комп'ютерний вчений, який, мабуть, найбільш відомий своєю багатотомною працею "Мистецтво комп'ютерного програмування". У цій роботі він присвячує цілий розділ темі генерації випадкових чисел і надає вичерпний огляд цієї галузі. Він також розробив кілька ГПСЧ, включаючи генератори MIXMAX і MMIX.

Марсалья та Заман: Марсалья та Заман розробили статистичний тест для оцінки якості ГПСЧ, який називається "спектральний тест". Цей тест досліджує частотний спектр згенерованої послідовності і може виявити певні типи патернів, які свідчать про низьку якість ГПСЧ.

П'єр Л'Екуйє: Л'Екуйє - математик і комп'ютерний науковець, який зробив багато внесків

у сферу генерації випадкових чисел. Він розробив кілька широко використовуваних ГПСЧ, включаючи генератори WELL і LFSR. Він також розробив кілька статистичних тестів для оцінки якості генераторів випадкових чисел, включаючи набір тестів TestU01.

Ці та багато інших вчених зробили значний внесок в область генерації випадкових чисел і статистичного аналізу випадковості послідовностей. Їхні роботи допомогли покращити якість ГПВЧ і дозволила більш ефективно використовувати їх у широкому спектрі застосувань, включаючи IoT.

3. Мета і задачі дослідження.

Загалом, ГПСЧ відіграють вирішальну роль у забезпеченні безпеки, випадковість і стійкість до систем IoT. Однак важливо використовувати високоякісний алгоритм ГПСЧ, щоб гарантувати, що значення, які генеруються, є дійсно випадковими та безпечними, а також належним чином завантажувати дані для ГПСЧ, щоб запобігти передбачуваності його результату. Крім того, важливо регулярно оновлювати алгоритми ГПСЧ і початкові значення, щоб гарантувати, що згенеровані значення залишатимуться безпечними з плином часу.

Саме тому метою роботи є проведення статистичного аналізу випадковості послідовностей, що породжуються генераторами випадкових та псевдовипадкових послідовностей на основі методів програмної інженерії.

4. Результати дослідження.

Мінімальний розмір вибірки обмежений не стільки алгоритмом розрахунку критерію, скільки розподілом його статистики. Таким чином, для ряду алгоритмів із занадто малими числами нормальна апроксимація розподілу статистики критерію буде під питанням. Використаємо класифікацію, наведену в [7]. У цьому випадку досліджувану пропускну здатність пам'яті важко піддається корекції, перевіряючи за допомогою одновимірної статистики. У цій ситуації ми пропонуємо перевірити послідовність на випадковість за допомогою дво- та/або тривимірної статистики. Опишемо відповідну математичну модель.

Спільний розподіл кількості 2-ланцюгів і кількості 3-ланцюгів заданого типу в двійковій послідовності

Нехай існує випадковий або псевдовипадковий процес і асоційовані з ним випадкові величини

$$\gamma_1, \gamma_2, \dots, \gamma_n, \tag{1}$$

де $\gamma_i = \{0, 1\}$, i = 1, 2, . . . , n , $n > 0$.

Підпослідовність $\gamma_j, \gamma_{j+1}, \dots, \gamma_{j+s-1}$, послідовності (1) називається s-ланцюгом $j = 1, 2, \dots, n - s + 1, s = 1, 2, \dots, n$.

Позначимо $\eta(t_1 , t_2, \dots, t_s$)кількість s-ланцюгів у послідовності (1), які збігаються з $t_1 , t_2, \dots, t_s$, де $t_i = \{0, 1\}, i = 1, 2, \dots, s$.

Умова (С): послідовність (1) складається з n , n >0, незалежних однаково розподілених випадкових величин;$P\{\gamma_i = 1\} = p$ $P\{\gamma_i = 0\} = q$, $p + q$ = 1, i = 1, 2, . . . , n .

Теорема. Нехай виконуються умови (С), k_1, k_2, k_3, t, t_1– цілі числа, що $k_1 \geq 0, k_2 \geq 0, n \geq k_1, k_3 \geq 0, t, t_1 \epsilon \{0, 1\}$. Потім:

$$P\{\eta(t, t^*) = k_1, \eta(t\ 1\ t^*) + \eta(t\ 0\ t^*) = k_2\ \} = \sum_{m_1=k}^{n-k_1} p^{m_1} q^{m_0} \times \sum \prod_{i=0}^{1} C_{k_1}^{\delta_i} C_{m_i-k_1}^{k_1-\delta_i}, \tag{2}$$

де $m_0 = n - m_1$, символ $\sum$ позначає додавання над усіма невід'ємними цілими числами δ_0і δ_1такими, що $\delta_0 + \delta_1 = 2k_1 - k_2$, $t^* \stackrel{\text{def}}{=} 1 - t$;

$$P\{\eta(t_1, t_1^*) = k_1, \eta(t_1\ t\ t_1^*) = k_2\ \} = \sum_{m_1=k}^{n-k_1} p^{m_1} q^{m_0}\ C_{k_1}^{k_2} C_{m_1-k_1}^{k_2} C_{m_{t^*}}^{k_1}, \tag{3}$$

$$P\{\eta(t_1, t_1^*) = k_1, \eta(t_1\ 1\ t_1^*) = k_2, \quad \eta(t_1\ 0\ t_1^*) = k_3\ \} = \sum_{m_1=k}^{n-k_1} p^{m_1} q^{m_0}\ C_{k_1}^{k_2} C_{k_1}^{k_3} C_{m_{t_1-k_1}}^{k_2} C_{m_{t_1^*}-k_1}^{k_3}. \tag{4}$$

Математично-статистичний аналіз послідовностей, як правило, відбувається в два етапи. Схематично процес аналізу послідовностей зображено на рис. 1.

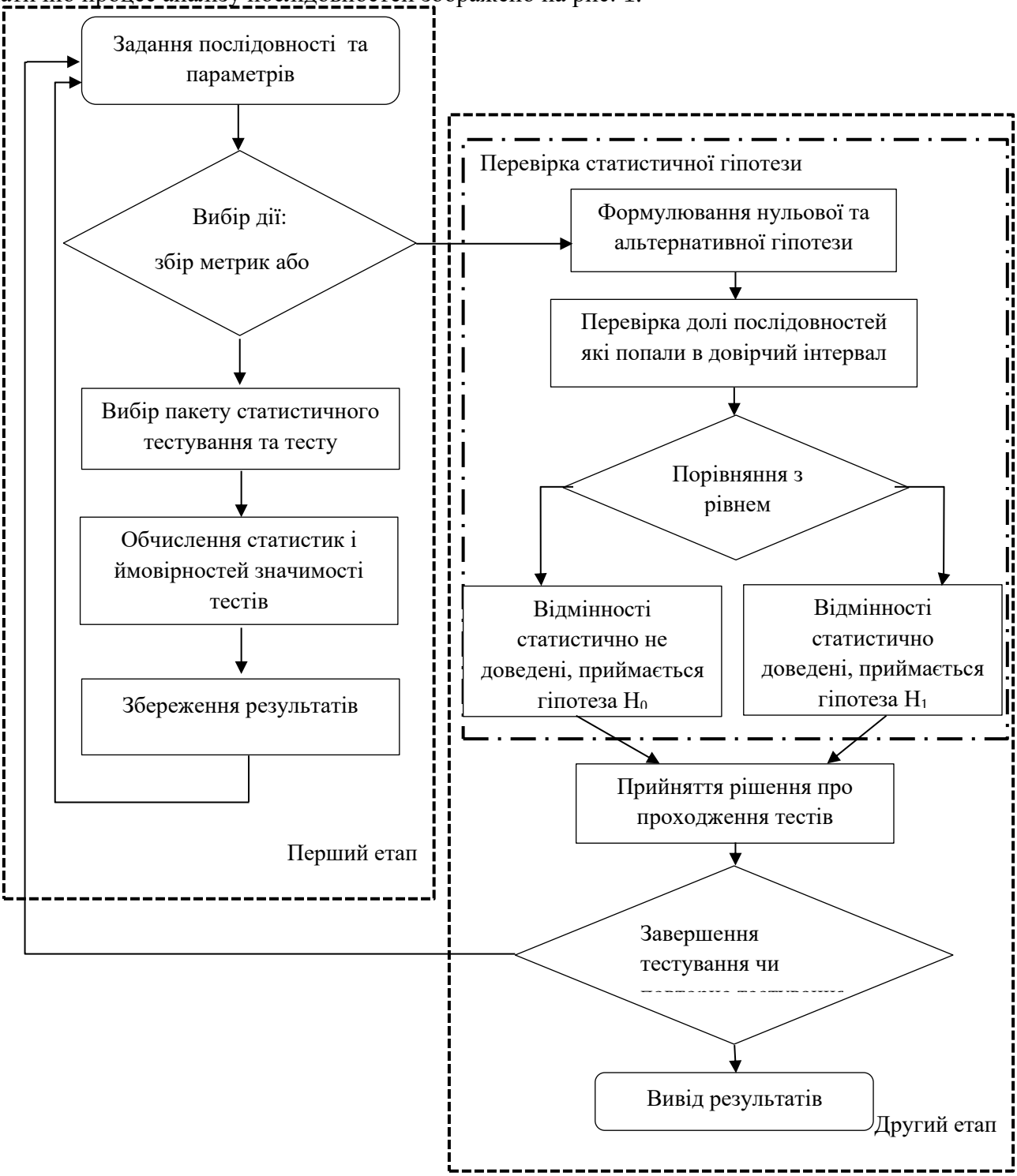

Рис. 1. Схема статистичного аналізу послідовностей

Використовуючи основи програмної інженерії, зокрема, способи нотації діаграм, побудуємо контекстну діаграму (IDEF0) для системи тестування послідовностей на випадковість (рис.2).

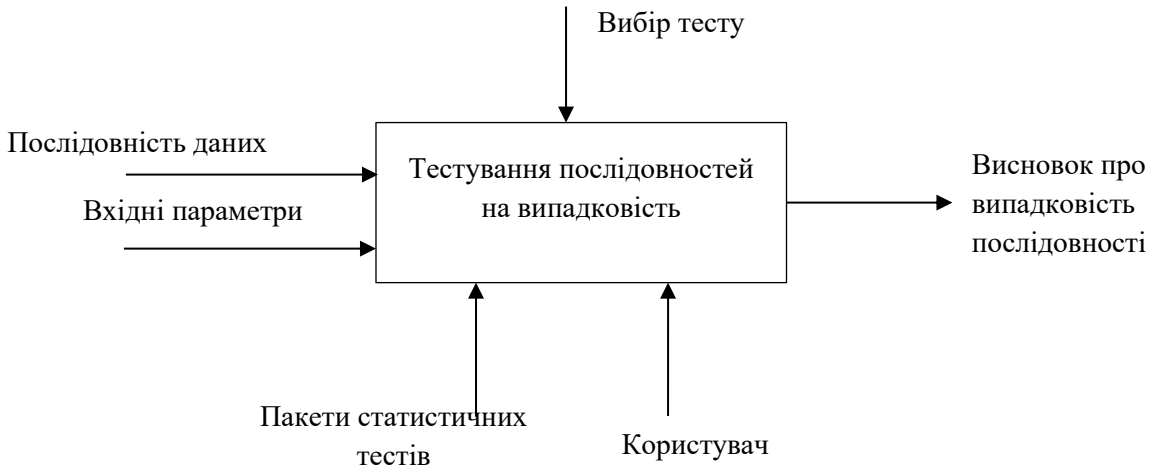

Рис 2. Контекстна діаграма

Приклади використання двійкової статистики для аналізу бітової послідовності

Розглянемо послідовності довжиною n = 12.

Таблиця 1

Таблиця ймовірностей послідовностей довжиною 12 для формули (2)

(k_1, k_2)	П	P_c
(3, 6)	0,000244	0,000244
(6, 0)	0,000244	0,000488
(5, 0)	0,000732	0,001221
(4, 0)	0,001221	0,002441
(3, 0)	0,001709	0,00415
(2, 0)	0,002197	0,006348
(1, 0)	0,002686	0,009033
(0, 0)	0,003174	0,012207
(5, 1)	0,007324	0,019531
(3, 5)	0,010254	0,029785
(5, 2)	0,010986	0,040771
(4, 4)	0,01709	0,057861
(4, 1)	0,019531	0,077393
(1, 1)	0,026855	0,104248
(2, 4)	0,030762	0,13501
(3, 1)	0,030762	0,165771
(2, 1)	0,035156	0,200928
(1, 2)	0,040283	0,241211
(4, 2)	0,068359	0,30957
(4, 3)	0,068359	0,37793
(3, 4)	0,076904	0,454834
(2, 2)	0,123047	0,577881
(2, 3)	0,123047	0,700928
(3, 2)	0,128174	0,829102
(3, 3)	0,170898	1

Побудувавши множини всіх можливих ланцюжків довжиною 12 елементів, отримаємо кількість ланцюгів, що дорівнює 2^{12}.

Проаналізуємо твердження (2). Для цього фіксуємо, наприклад, $t = 0$ і обчислюємо

кількість k_1 і k_2 для кожної з можливих послідовностей довжиною 12, де – k_1 кількість входжень підпослідовності типу (01), а k_2 – кількість входжень підпослідовності типу (001) або (011) відповідно.

Таблиця 1 і рис. 3 показано використання співвідношення (2) для вибірки n , n = 12, і деяких значень k_1 і k_2 .

Таблиця 2 у першому стовпчику містить усі можливі варіанти k_1 і k_2 , для яких імовірність $P\{\eta(t\ t^*) = k_1, \eta(t\ 1\ t^*) + \eta(t\ 0\ t^*) = k_2\} \geq 0{,}0001$.У другому стовпчику таблиці 2 подано ймовірності (у порядку неубування)$P\{\eta(t\ t^*) = k_1, \eta(t\ 1\ t^*) + \eta(t\ 0\ t^*) = k_2\}$ для пар чисел(k_1, k_2), зазначених у першій колонці.

Кожен рядок третього стовпця містить суму накопичених ймовірностей до реалізації події $\{\eta(t\ t^*) = k_1, \eta(t\ 1\ t^*) + \eta(t\ 0\ t^*) = k_2\}$ включно, де k_1 і k_2 вказані в цьому ж рядку першого стовпця.

На рис. 3 показано ймовірність$P\{\eta(t\ t^*) = k_1, \eta(t\ 1\ t^*) + \eta(t\ 0\ t^*) = k_2\} \geq 0{,}01$.

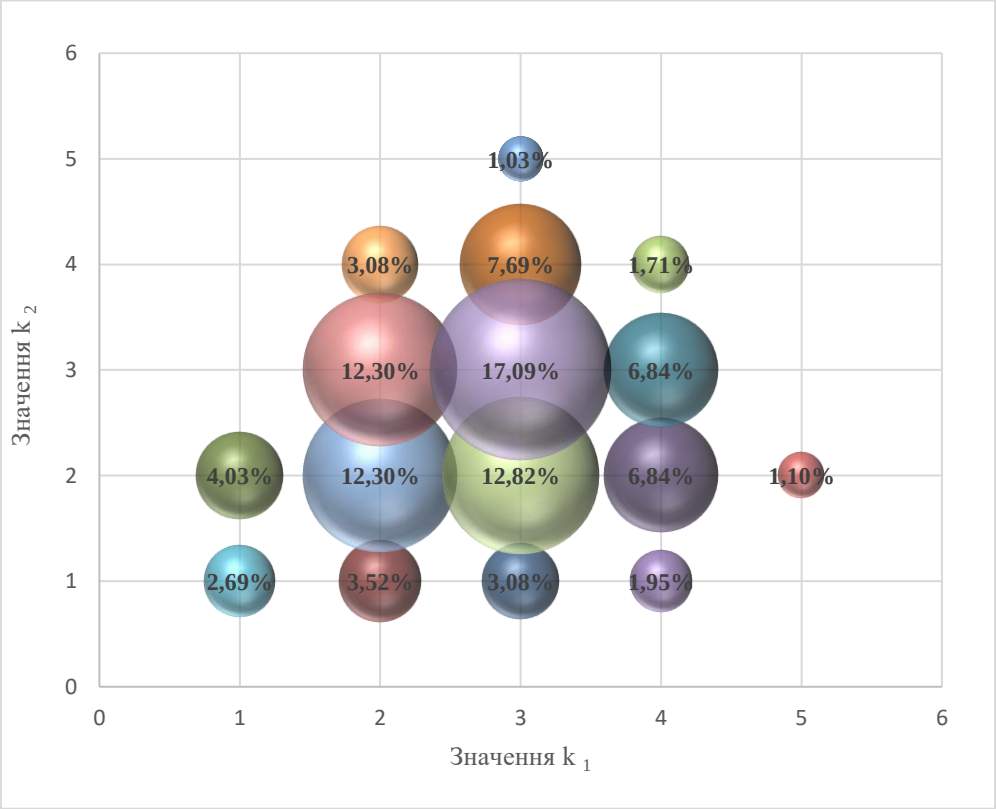

Рис. 3. Розподіл ймовірностей залежно від k_1 і k_2 за формулою (2) для послідовностей довжиною 12 .

Проаналізувавши рисунок 1, можна зробити висновок, що для аналізу послідовності ланцюжків дуже малої довжини (від 5 до 12 елементів) [7] одновимірна статистика не завжди дає правильний результат. Наприклад, якщо ми розглядаємо послідовність (0 0 1 0 0 1 1 1 0 0 1 1), де $k_1 = 3$, то з високою ймовірністю можна зробити висновок, що послідовність з цими характеристиками є випадковою, але якщо звернути увагу, коли $k_1 = 3$ і $k_2 = 5$, можна стверджувати, що ця послідовність невипадкова. Це свідчить про відсутність використання одновимірної статистики для аналізу коротких і середніх бітових послідовностей.

Розглянемо послідовності довжиною n =10.

Побудувавши множини всіх можливих ланцюгів довжиною 10 елементів, отримаємо кількість ланцюгів, що дорівнює 2^{10} .

Проаналізуємо твердження (3). Зафіксуйте t =0 і t_1 =1. Для цього обчислюємо кількість

k_1 і k_2, для кожного з можливі послідовності довжиною 10, де k_1 це кількість входжень підпослідовності типу (01) і k_2 кількість входжень підпослідовності типу (001).

Таблиця 2 і рис. 4 показано використання співвідношення (3) для вибірки n , n = 10, і деяких значень k_1 і k_2.

Таблиця 2 у першому стовпчику містить усі можливі варіанти k_1 і k_2, для яких імовірність $P\{\eta(t_1, t_1^*) = k_1, \eta(t_1\ t\ t_1^*) = k_2\} \geq 0{,}0001$. У другому стовпчику таблиці 2 подано ймовірності (у порядку неубування) $P\{\eta(t_1, t_1^*) = k_1, \eta(t_1\ t\ t_1^*) = k_2\}$ для пар чисел (k_1, k_2), зазначених у першій колонці.

Таблиця 2

Таблиця ймовірностей послідовностей довжиною 10 для формули (3)

(k_1, k_2)	П	P_c
(3, 3)	0,001953	0,001953
(4, 1)	0,007813	0,009766
(4, 0)	0,011719	0,021484
(0, 0)	0,019531	0,041016
(3, 2)	0,041016	0,082031
(3, 0)	0,068359	0,150391
(1, 0)	0,070313	0,220703
(2, 0)	0,109375	0,330078
(2, 2)	0,109375	0,439453
(3, 1)	0,123047	0,5625
(1, 1)	0,164063	0,726563
(2, 1)	0,273438	1

Кожен рядок третього стовпця містить суму накопичених ймовірностей до реалізації події $\{\eta(t_1, t_1^*) = k_1, \eta(t_1\ t\ t_1^*) = k_2\}$ включно, де k_1 і k_2 вказані в цьому ж рядку першого стовпця.

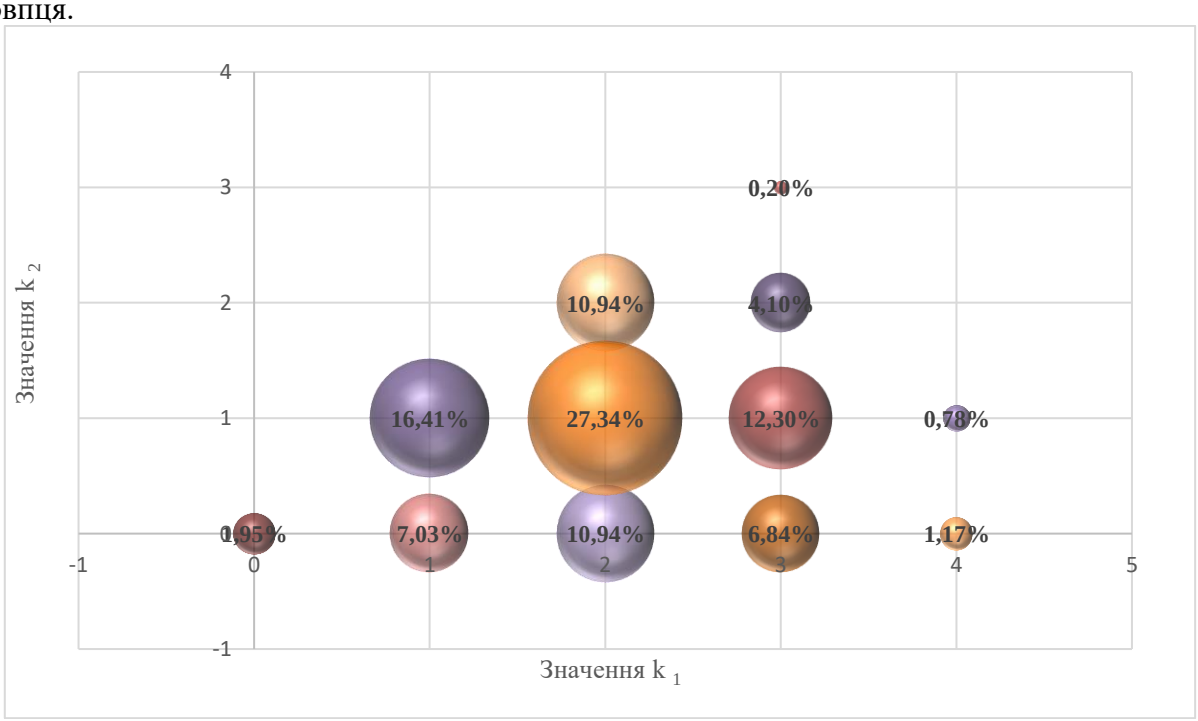

Рис. 4. Розподіл ймовірностей залежно від k1 і k2 за формулою (3) для послідовностей довжиною 10 .

На рис. 4 показано ймовірність$P\{\eta(t_1, t_1^*) = k_1, \eta(t_1\ t\ t_1^*) = k_2\} \geq 0{,}001$.

5. Висновки і перспективи подальших досліджень.

Використання статистичного аналізу випадковості послідовностей, що генеруються генераторами випадкових (або псевдовипадкових) чисел, є популярною тенденцією, оскільки це має вирішальне значення для забезпечення надійності та безпеки різних комп'ютерних систем і додатків, включаючи Інтернет речей (IoT), криптографію, моделювання та наукові обчислення. Випадковість є невід'ємною властивістю багатьох додатків, а якість генераторів випадкових чисел безпосередньо впливає на безпеку, точність та ефективність цих додатків. Тому важливо оцінювати випадковість послідовностей, згенерованих генераторами випадкових (або псевдовипадкових) чисел, за допомогою статистичних тестів, щоб переконатися, що вони відповідають бажаним стандартам випадковості. Зі зростанням важливості кібербезпеки та конфіденційності даних попит на надійні та безпечні генератори випадкових чисел зростає, а використання статистичного аналізу для оцінки випадковості послідовностей, згенерованих генераторами випадкових (або псевдовипадкових) чисел, стає все більш критичним.

Список використаних джерел:

1. Bhaskar C. U., Rupa C. An advanced symmetric block cipher based on chaotic systems. In: The 2017 Innovations in Power and Advanced Computing Technologies (i-PACT), pp. 1-4. IEEE Press, Vellore (2017)

2. Busireddygari P.; Kak S. Pseudorandom tableau sequences, In: 51st Asilomar Conference on Signals, Systems, and Computers, pp. 1733 – 1736. IEEE Press (2017)

3. Gurugopinath S., Samudhyatha B., Multi-dimensional Anderson-Darling statistic based goodness-of-fit test for spectrum sensing. In: Seventh International Workshop on Signal Design and its Applications in Communications (IWSDA). pp. 165-169. Bengaluru, India. (2015).

4. Moody D. Post-quantum cryptography: NIST's plan for the future. In: Proceedings of the Seventh International Conference on Post Quantum Cryptography. IEEE Press, Japan, (2016). https://pqcrypto2016.jp

5. Nejad F. H., Sabah S., Jam A. J. Analysis of avalanche effect on advance encryption standard by using dynamic S-Box depends on rounds keys. In: The 2014 International Conference on Computational Science and Technology (ICCST), pp. 1-5. IEEE Press, Kota Kinabalu (2014)

6. Popereshnyak S. Analysis of pseudorandom small sequences using multidimensional statistics. In: The 3rd IEEE International Conference on Advanced Information and Communication Technologies (AICT), pp. 5.4.1-5.4.4. IEEE Press, Ukraine (2019)

7. Special Publication 800-22. A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications. http://csrc. nist.gov

8. V Masol, S Popereshnyak A theorem on the distribution of the rank of a sparse Boolean random matrix and some applications // Theory of Probability and Mathematical Statistics/ – 2008. – V. 76. – P. 103-116

9. Yueqi Hu; Tom Polk; Jing Yang; Ye Zhao; Shixia Liu Spot-tracking lens: A zoomable user interface for animated bubble charts // IEEE Pacific Visualization Symposium (PacificVis) – 2016. – pp. 16-23

10. Вимоги безпеки до криптографічних модулів [Електронний ресурс]. Режим доступу: http://csrc.nist.gov/publications/fips/fips1402/fips1402.pdf

11. Гайдишев І. П. Програмне забезпечення аналізу даних AtteStat. Руководство пользователя. Версія 13 – 2012. – 505 с.

12. Коренєва А.М., Фомічов В.М. «Статистичне тестування псевдослучайных последствий» // Безопасность информационных технологий № 2 2016 г .

References:

1. 1. Bhaskar C.U., Rupa C. An advanced symmetric block cipher based on chaotic systems. In: The 2017 Innovations in Power and Advanced Computing Technologies (i-PACT), pp. 1-4. IEEE Press, Vellore (2017)

2. Busireddygari P.; Kak S. Pseudorandom tableau sequences, In: 51st Asilomar Conference on Signals, Systems, and Computers, pp. 1733 – 1736. IEEE Press (2017)

3. Gurugopinath S., Samudhyatha B., Multi-dimensional Anderson-Darling statistic based goodness-of-fit test for spectrum sensing. In: Seventh International Workshop on Signal Design and its Applications in Communications (IWSDA). pp. 165-169. Bengaluru, India. (2015).

4. Moody D. Post-quantum cryptography: NIST's plan for the future. In: Proceedings of the Seventh International Conference on Post Quantum Cryptography. IEEE Press, Japan, (2016). https://pqcrypto2016.jp

5. Nejad F.H., Sabah S., Jam A.J. Analysis of avalanche effect on advance encryption standard by using dynamic S-Box depends on round keys. In: The 2014 International Conference on Computational Science and Technology (ICCST), pp. 1-5. IEEE Press, Kota Kinabalu (2014)

6. Popereshnyak S. Analysis of pseudorandom small sequences using multidimensional statistics. In: The 3rd IEEE International Conference on Advanced Information and Communication Technologies (AICT), pp. 5.4.1-5.4.4. IEEE Press, Ukraine (2019)

7. Special Publication 800-22. A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications. http://csrc. nist.gov

8. V Masol, S Popereshnyak A theorem on the distribution of the rank of a sparse Boolean random matrix and some applications // Theory of Probability and Mathematical Statistics/ – 2008. – V. 76. – P. 103-116

9. Yueqi Hu; Tom Polk; Jing Yang; Ye Zhao; Shixia Liu Spot-tracking lens: A zoomable user interface for animated bubble charts // IEEE Pacific Visualization Symposium (PacificVis) - 2016. - 16-23

10. Security requirements for cryptographic modules [Electronic resource]. Access mode: http://csrc.nist.gov/publications/fips/fips1402/fips1402.pdf

11. Gaidyshev I.P. AtteStat data analysis software. Руководство пользователя. Version 13 – 2012. – 505 p.

12. Koreneva A.M., Fomichev V.M. "Statistical testing of pseudo-random effects" // Security of information technologies No. 2, 2016.