computing@computingonline.net
www.computingonline.net

Print ISSN 1727-6209
On-line ISSN 2312-5381
International Journal of Computing

TECHNIQUE OF THE TESTING OF PSEUDORANDOM SEQUENCES

Svitlana Popereshnyak

Taras Shevchenko National University of Kyiv, 24, Bohdana Havrylyshyna str., Kyiv, 04116, Ukraine,
spopereshnyak@gmail.com

Paper history:
Received 30 July 2019
Received in revised form 02 March 2020
Accepted 08 July 2020…
Available online 27 September 2020

Keywords:
algorithms;
multidimensional statistics;
random sequence;
s-chains;
cryptography;
pseudorandom sequence;
statistical testing.

Abstract: The article is dedicated to systematization of scientific positions about the static testing of sequences, widely used in cryptographic systems of information protection for the production of key and additional information (random numbers, vectors of initialization, etc.). Existing approaches to testing pseudorandom sequences, their advantages and disadvantages are considered. It is revealed that for sequences of length up to 100 bits there are not enough existing statistical packets. Perspective direction of research – static testing of sequences using n- dimensional statistics is considered. The joint distributions of 2-chains and 3-chains of a fixed type of random (0, 1) -sequences allow for statistical analysis of local sections of this sequence. Examples, tables, diagrams that can be used to test for randomness of the location of zeros and ones in the bit section are 16 lengths. The paper proposes a methodology for testing pseudorandom sequences, an explicit form of the joint distribution of 2- and 3-chains numbers of various options of random bit sequence of a given small length is obtained. As a result of the implementation of this technique, an information system will be created that will allow analyzing the pseudorandom sequence of a small length and choosing a quality pseudorandom sequence for use in a particular subject area.

1. INTRODUCTION

Random sequences have found the widest application from the gaming computer industry to mathematical modeling and cryptology.

We list some areas of their usage:

1. Modeling. In computer simulation of physical phenomena. In addition, mathematical modeling uses random numbers as one of the tools of numerical analysis.

2. Cryptography and information security. Random numbers can be used to test the correctness or effectiveness of algorithms and programs. Many algorithms use the generation of pseudo-random numbers to solve applied problems (for example, cryptographic encryption algorithms, the generation of unique identifiers, etc.).

3. Decision making in automated expert systems. The use of random numbers is part of decision-making strategies. For example, for the impartiality of the choice of examination paper by a student in an exam. Randomness is also used in the theory of matrix games.

4. Optimization of functional dependencies. Some mathematical optimization methods use stochastic methods to search for extremums of functions.

5. Fun and games. Accident in games has a significant role. In computer or board games, chance helps to diversify the gameplay.

There are various approaches to the formal definition of the term “randomness” based on the concepts of computability and algorithmic complexity [1].

By implementing some algorithm, software generators produce numbers (although not obvious) depending on the set of previous values, so the received numerical sequences are not truly random and are called pseudo-random sequences (PRS). At the moment, more than a thousand software PRS generators are known, which differ in algorithms and values of parameters. Statistical properties are

significantly different from the number sequences that are generated by them.

The presented and not presented results allow us to characterize the state of modern technologies of designing the PRS (focusing on the most progressive of them by the following basic provisions [2-13].

2. REVIEW OF EXISTING SETS OF PRS TESTS AND THEIR APPLICATION

A selection of 14 tests "Diehard" J. Marsaly was the first in the complex testing of generators PRS. The selection is considered as one of the most rigorous test suites; implemented software and available on the Internet. However, the selection of tests "Diehard" has several disadvantages.

– There is no detailed description of the tests and methods for interpreting the results.
– Test parameters are hard coded. At the same time, regardless of the length of the PRS being tested, only a certain number of bytes is analyzed. Shorter PRS cannot be tested.
– Most of the tests are heuristic and based on test results, rather than on theoretical models.
– The decision to pass the test can take only one of two values (yes / no).

Compilation of tests PRS. D. Knut uses seven original statistics and algorithms for their calculation. However, this collection has several disadvantages.

– All algorithms are reduced to the calculation of statistical criteria that are approximated only by the distribution of χ^2.
– There are no recommendations for test parameters. Incorrect selection of some values can lead to a significant dependence on the length of the tested sequence, as well as adversely affect the power of the statistical criterion.
– A controversial approach [2] is the method of evaluating results, when sequences are recognized as random, for which the P- value belongs to the interval (0,1; 0,9). That is, when P-value > 0,9, the test results are considered too ideal to consider a numerical sequence to be random.
– There is no original software implementation of the proposed tests.

A set of tests is proposed for preliminary testing of the quality of random numbers and sequences based on seven different statistical tests.

Kendal M. and Smith B. suggested performing 4 tests using χ^2 test:

– checking the frequency of different digits x_1, x_2, ..., x_N in the table (frequency test);
– checking the frequency of different two-digit numbers among pairs of digits x_1x_2, x_2x_3, x_3x_4, ..., $x_{N-1}x_N$ (test pairs);
– checking the frequency of different intervals between two consecutive zeros (test intervals);
– checking the frequency of different types of quadruples (*aaaa*, *aaab*, *aabc*, *aabb*, *abcd*) among quadruples $x_1x_2x_3x_4$, $x_2x_3x_4x_5$, $x_3x_4x_5x_6$, ..., $x_{N-3}x_{N-2}x_{N-1}x_N$; and also checking the frequency of various types of the fives (pokertest).

The NIST STS 800-22 standard of the National Institute of Standardization and Technologies NIST [14] includes 15 tests and is focused on testing bit sequences used in the tasks of cryptographic protection of information.

A typical application of tests (in particular, Diehard) is given, for example, in the report.

With an increase in the length of the tested memory bandwidth (more than 100 thousand), many statistical tests begin to detect statistically significant patterns that were not found on samples of smaller size. For example, the sign rank criterion (signed rank test, Wilcoxon), which is quite powerful, rejects such well-known and high-quality generators, as Bluma-Blum-Shuba (BBS), Shamir (RSA), "Marsaglia Multicarry" and "Xorshift" George Marsala Mersenne vortex (MT19937), as well as "truly random sequence" having 1.5-2 thousands of elements of a numerical sequence.

The stream encryption of a long sequence has the most significant potential advantage over block cryptographic transformations [15-22], which is essential for many applications [23-27].

Dimensionality reduction without losing essential information is the goal of any approach designed to cope with high-dimensional time sequences. In this relation, [28-29] should be mentioned first of all. It enables evaluation of distance between any two-sample series from a sequence of observations.

The results obtained in the paper [30] are applied to estimate the probability that a nonhomogeneous system of Boolean random linear equations is consistent.

An overview of popular methods for testing bit sequences for randomness showed that, despite the large number of statistical tests, they all give a more correct result with a sufficiently large sample size. However, we will not be able to get a correct answer about the randomness of the sequence if the sequence length is less than 100 elements. In this situation, we propose to test the sequence for randomness using two and / or three-dimensional statistics.

3. PROBLEM STATEMENT

Before responsible using in mathematical modeling and cryptology, PRS should be tested. Unfortunately, for many PRS tests, there are some limitations:

– check out only one of the probable ones properties that characterize PRS;
– do not fix family alternatives;
– do not have theoretical one's ratings power.
– do not give a correct estimate of chance sequences providing a little sample.

Problems of small and large samples refer to the main problems that arise in practical application methods of data analysis. Let us use the next classification samples by number [31], based on requirements presented in the program criteria:

– very small sampling – from 5 to 12,
– small sampling – from 13 to 40,
– average sampling number – from 41 to 100,
– large sampling – from 101 and more.

The minimum size of the sample limits not so much the algorithm of calculating the criterion, but the distribution of its statistics. So, for row algorithms with too much small ones numbers sample normal approximation distribution of statistics criterion will be under question.

During the research, the localization of the local sections of the bit sequence was conducted to detect the dependencies in the location of its elements by using the exact distributions of the corresponding statistics. In the work an explicit form of the joint distribution of the numbers of 2-chains and numbers of 3-chains of various variants in a random sequence was obtained. This joint distribution allows more accurate comparison of the use of one-dimensional statistics, to analyze the bit sequence small length by chance [32-34].

4. MATERIALS AND METHODS

4.1. SCHEME FOR VERIFICATION OF STATISTICAL TESTS OF RANDOMNESS SEQUENCES

If users of mathematical and statistical algorithms and their software products are interested in quality research, the following steps should be performed before conducting any research (Fig. 1):

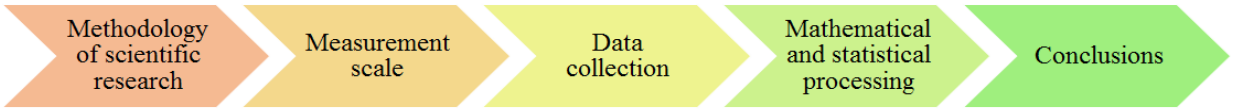

Figure 1 – Life cycle research

1. Examine philosophical bases of the methodology of scientific study.

2. Develop a clear understanding about scales measurement. It is through the scale measurement of the original data, methods that can be used for their processing are determined, in order to determine which method to use to help names modules software provision and their descriptions. Before applying of each method one should get acquainted with it prerequisites and constraints and plan necessary amount sampling based on power criteria.

3. Start collecting data. Already selected processing method asks in which form should be presented experimental results. Data can be adequately used by the predicted method.

4. Mathematical and statistical processing is penultimate, technical, stage, whose content should be completely understandable after implementation of the 2nd stage, while there was still no significant cost for the experimental study. This stage does not have any relation to the subject matter of the area.

5. The last one stage is objective scientifically justified conclusion based on the results of the study, taking into account subject matter industry, recommendations and forecast. Using the methods of chart notation, we construct a context diagram (IDEF 0) for the random sequence testing system (Fig. 2).

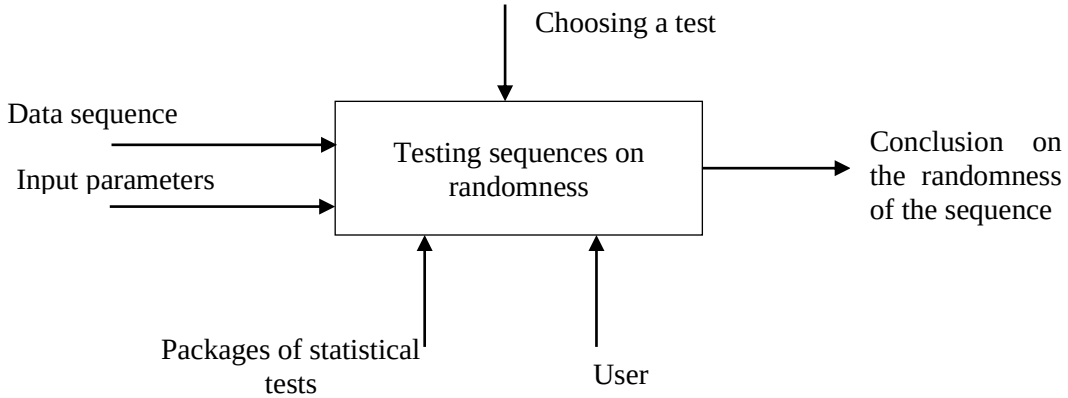

Figure 2 – Context diagram

Mathematical and statistical analysis of sequences, as a rule, takes place in two stages. Schematically the process of sequences analysis is depicted in Fig. 3.

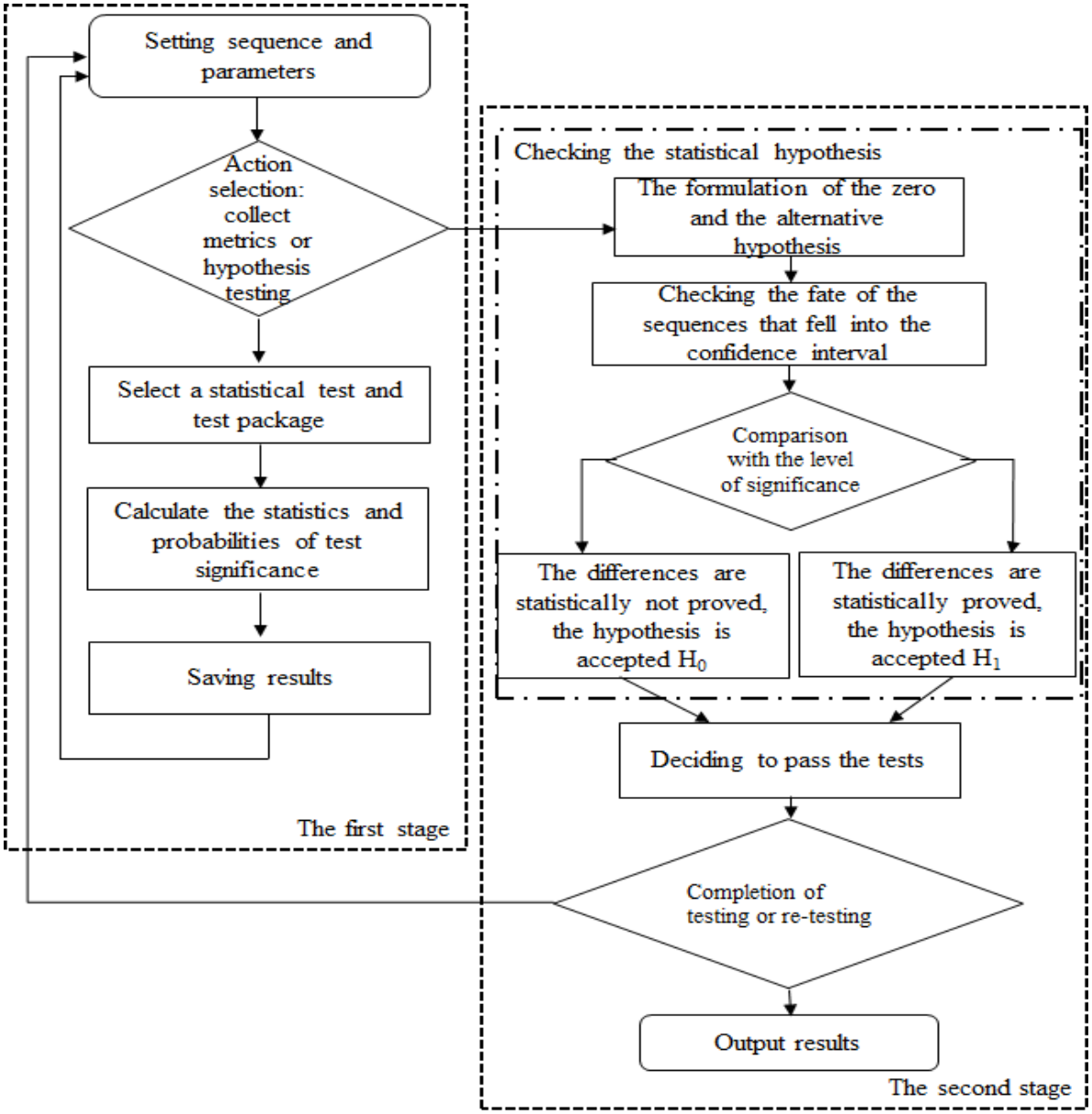

Figure 3 – Scheme of statistical analysis of sequences

Description of the main steps:

1. The first stage is named as preparatory, it is the most labor-intensive step, and here basic mass calculations are executed.

1.1. With the help of experimental generator casual sequences are formed (or given sequences are introduced).

1.2. For each sequence test statistics is calculated. If a battery of tests runs (conducted immediately several tests), then the statistics on the results is issued for each test.

1.3. Probability significance is calculated for each sequence.

1.4. Obtained statistics and probabilities significance are stored.

2. The second stage processes the received results.

2.1. Audit of statistical hypothesis.

2.1.1. Formulation of zero and alternative hypothesis.

2.1.2. With the help of criteria coordination, the hypothesis compliance of distributed statistical data and probabilities of meaningful hypothetical distributions is checked out.

2.1.3. Number sequences that passed the test are determined. It is being built trustworthy interval for the last of magnitude.

2.1.4. Comparison of fate sequences which are in the trust interval with level significance and acceptance decision on passing tests.

Trust probability is necessary to calculate a number of sample statistical indicators as well

differences from a number of others parameters that are not calculated by sampling, but are asked by user program size. It is selected from the following standard rulers:

– Zero threshold of 0,90 applies to work with lowered responsibility at the first familiarity with the phenomenon;

– The first threshold of 0,95 is applied in most studies (e.g., biological research);

– Second threshold of 0,99 is used to work with higher liability (e.g., medical research);

– Third threshold of 0,999 is used to work with highest liability (e.g., research efficiency medicine).

2.2. Decision is made whether you can consider the test to be passed.

2.3. If the results are satisfactory the decision will be made to complete the test, otherwise, go to step 1.2.

2.4. Final conclusions.

4.2. JOINT DISTRIBUTIONS OF THE NUMBER OF 2-CHAINS AND 3-CHAINS

Consider a sequence of random variables

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

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

Subsequences $\gamma_j, \gamma_{j+1}, \dots, \gamma_{j+s-1}$, sequences (1) are called s-chains, $j = 1, 2, \dots, n-s+1, s = 1, 2, \dots, n$.

Denote $\eta(t_1, t_2, \dots, t_s)$ the number of s-chains in the sequence (1) that coincide with $t_1, t_2, \dots, t_s$, where $t_i = \{0, 1\}$, $i = 1, 2, \dots, s$.

Theorem. Let sequence (1) consist of n, $n>0$, independent identically distributed random variables; $\mathrm{P}\{\gamma_i = 1\} = p$, $\mathrm{P}\{\gamma_i = 0\} = q$, $p + q = 1$, $i = 1, 2, \dots, n$ and $k_1, k_2, k_3, t,$ – integer numbers such that $k_1 \geq 0, k_2 \geq 0, n \geq k_1, k_3 \geq 0, t, t_1 \epsilon \{0, 1\}$. Then

$$\begin{gathered}\mathrm{P}\{\eta(t_1\, t_1^*) = k_1, \eta(\, t1t) + \eta(\, t0t) = k_2\,\} = \\ \sum_{m_1=k_1}^{m-k_1} p^{m_1} q^{m_0} \times \\ \sum\sum_{i\in\{k_1,\, k_1+1\}} C_{i-1}^{\delta_{t^*}} C_i^{\delta_t - m_t + 2i} C_{m_{t^*}-i+1}^{k_1-\delta_{t^*}} \times \\ \mathrm{Z}(m_t - i; m_t - i - \delta_t),\end{gathered} \tag{2}$$

where is the symbol $\sum$ denotes addition over all non-negative integers δ_t and δ_{t^*} such that $\delta_t + \delta_{t^*} = k_2$,

$$\mathrm{Z}(a, b) \stackrel{\text{def}}{=} \begin{cases} C_{a-1}^{b-1}, \text{ if } a \geq b \geq 1; \\ 1, \text{if } a = b = 0; \\ 0, \text{ elsewhere;} \end{cases}$$

$$\begin{gathered}\mathrm{P}\{\eta(t_1\, t_1^*) = k_1, \eta(\, ttt) = k_2\,\} = \\ \sum_{m_1=k_1}^{m-k_1} p^{m_1} q^{m_0}\, C_{m_{t^*}}^{k_1} \times \\ \sum_{i\in\{k_1,\, k_1+1\}} C_i^{m_t - k_2 - i}\, \mathrm{Z}(m_t - i, m_t - i - k_2);\end{gathered} \tag{3}$$

$$\begin{gathered}\mathrm{P}\{\eta(t_1\, t_1^*) = k_1, \eta(\, tt^*t) = k_2\,\} = \\ \sum_{m_1=k_1}^{m-k_1} p^{m_1} q^{m_0} \sum_{i\in\{k_1,\, k_1-1\}} C_i^{k_2}\, C_{m_{t^*}-i}^{k_1-k_2} \times \\ \mathrm{Z}(m_t;\, i+1);\end{gathered} \tag{4}$$

$$\begin{gathered}\mathrm{P}\{\eta(t_1\, t_1^*) = k_1, \eta(\, ttt) = k_2, \eta(\, tt^*t) = \\ k_3\,\} = \sum_{m_1=k_1}^{m-k_1} p^{m_1} q^{m_0} \times \\ \sum_{i\in\{k_1,\, k_1+1\}} C_i^{k_2 - m_t + 2i}\, C_{i-1}^{k_3} C_{m_{t^*}-i+1}^{k_1-k_3} \times \\ \mathrm{Z}(m_t - i;\, m_t - i - k_2).\end{gathered} \tag{5}$$

Proof:

Check (3). For this purpose, we denote by ν the number of units in a random sequence (1). The random variable ν has a binomial distribution with parameters (n, p). This makes it possible to write for m = 0,1,2, ..., n the probability of the event $\{\nu = m\}$, namely:

$$\mathrm{P}\{\nu = m\} = C_n^m p^m q^{n-m}. \tag{6}$$

Using the formula for the total probability we find

$$\begin{gathered}\mathrm{P}\{B_1, B_2\} = \sum_{m_1=0}^{n} \mathrm{P}\{\nu = m\} \cdot \\ \mathrm{P}\{A_1, A_2/\upsilon = m_1\},\end{gathered} \tag{7}$$

where $B_1 \stackrel{\text{def}}{=} \{\eta(t_1\, t_1^*) = k_1\}$, $B_2 \stackrel{\text{def}}{=} \{\eta(ttt) = k_2\}$.

Let Q denotes the number of all vectors $\vec{v}$, $\vec{v} \in \Omega(n, m_1)$, of which has k_1 /k_2/ 2-chain type $t_1 t_1^*$ / 3-chains ttt / type. Then, taking into account $\Omega(n, m_1) = C_n^{m_1}$ we have

$$\mathrm{P}\{B_1, B_2/\nu = m_1\} = \left(C_n^{m_1}\right)^{-1} Q. \tag{8}$$

Next, set the formula for finding the Q number. For this purpose, we consider a subset $\mathrm{D}(m_0, m_1, k_1, k_2;\, t) \subseteq \Omega(n, m_1)$ all vectors that start and end with element t, contain k_1 /k_2/ 2-chain type $t_1 t_1^*$ / 3-chain ttt / type. Note that a random $\vec{v} \in \mathrm{D}(m_0, m_1, k_1, k_2;\, t)$ vector has the following property: vector $\vec{v}$ permutation between themselves $\propto$ – series, $\propto \in \{0, 1\}$, does not change the number k_1 and k_2. This allows you to write equality

$$\begin{gathered}Q = \sum_{\nu_{t^*}=0}^{m_{t^*}} |\mathrm{D}(m_{t^*} - \nu_{t^*}, m_t, k_1, k_2;\, t)| + \\ \sum_{\nu_{t^*}=0}^{m_{t^*}} \sum_{\nu'_{t^*}=1}^{m_{t^*}} \left|\mathrm{D}(m_{t^*} - \nu_{t^*} - \nu'_{t^*}, m_t, k_1 - \right. \\ \left. 1, k_2;\, t)\right|\end{gathered} \tag{9}$$

Show that

$$\begin{gathered}\mathrm{D}(m_0, m_1, k_1, k_2;\, t) = \\ C_{k_1+1}^{m_t - k_2 - k_1 - 1} C_{m_{t^*}-1}^{k_1-1} C_{m_t - k_1 - 2}^{k_2}.\end{gathered} \tag{10}$$

Indeed, for an arbitrary vector $\vec{v} \in \mathrm{D}(m_0, m_1, k_1, k_2;\, t)$ we have

$$k_2 = m_t - \delta_1^{(t)} - 2\left(k_1 - \delta_1^{(t)} + 1\right) = \delta_1^{(t)} + m_t - 2k_1 - 2, \quad (11)$$

where $\delta_1^{(t)} + m_t - 2k_1 - 2 \geq 0$, $\delta_1^{(t)}$ – number t-series unit length in vector $\vec{v}$. Element from set $\mathrm{D}(m_0, m_1, k_1, k_2;\ t)$ is uniquely determined if fixed as one of $C_{k_1+1}^{\delta_1^{(t)}}$ possible locations t-series of unit length each, as well as one of possible splits $m_{t^*} = x_1^{(t^*)} + x_2^{(t^*)} + \cdots + x_{k_1}^{(t^*)}$ on k_1 t-series, the length of each of which is not less than two. From here we have

$$|\mathrm{D}(m_0, m_1, k_1, k_2;\ t_1)| = C_{k_1+1}^{\delta_1^{(t)}} C_{m_{t^*}-1}^{k_1-1} C_{m_{t-k_1-2}}^{k_1-\delta_1^{(t)}} = C_{k_1+1}^{m_t-k_2-k_1-1} C_{m_{t^*}-1}^{k_1-1} C_{m_t-k_1-2}^{k_2}.$$

With the help of (10) and (11) we find

$$\sum_{\nu_{t^*}=0}^{m_{t^*}} |\mathrm{D}(m_{t^*} - \nu_{t^*}, m_t, k_1, k_2;\ t)| = C_{k_1+1}^{m_t-k_2-k_1-1} C_{m_{t^*}}^{k_1} C_{m_t-k_1-2}^{k_2} \sum_{\nu_{t^*}=0}^{m_{t^*}} \sum_{\nu'_{t^*}=1}^{m_{t^*}} |\mathrm{D}(m_{t^*} - \nu_{t^*} - \nu'_{t^*}, m_t, k_1 - 1, k_2;\ t)| = C_{k_1}^{m_t-k_2-k_1} C_{m_t-k_1-1}^{k_2} \sum_{\nu'_{t^*}=1}^{m_{t^*}} C_{m_{t^*}-\nu'_{t^*}}^{k_1-1}.$$

In this way,

$$Q = C_{m_{t^*}}^{k_1}\left(C_{k_1+1}^{m_t-k_2-k_1-1} C_{m_t-k_1-2}^{k_2} + C_{k_1}^{m_t-k_2-k_1} C_{m_t-k_1-1}^{k_2}\right),$$

that together with (6) – (9) prove (3).

5. EXPERIMENT

As a result of the application of this technique for testing pseudo-random sequences, tables were constructed, with the help of which one can obtain the probability of the distribution of zeros and ones in a given sequence. As practice shows, the use of ready-made tables for analyzing the sequence of randomness allows you to get the answer as quickly as possible, in contrast to the classical testing method.

Consider an example of tables for a bit-sequence of small length. For example, let the length of the bit sequence *n*, *n*=16.

5.1. ILLUSTRATION OF THE USE OF EQUALITY (2)

Table 1 and Fig. 4 show the use of the relation (2) for a small sample n, $n = 16$, and some values k_1 and k_2.

Table 1. Using (2) for a small sample of length 16

k_1	k_2	P	P_c
2	5	0,010910034	0,114364624
2	1	0,011276245	0,125640869
5	1	0,011291504	0,136932373
6	4	0,011672974	0,148605347
2	4	0,012207031	0,160812378
2	2	0,012252808	0,173065186
2	3	0,012695313	0,185760498
4	0	0,012985229	0,198745728
4	7	0,013580322	0,21232605
5	6	0,014358521	0,22668457
3	7	0,01550293	0,2421875
3	0	0,021987915	0,264175415
3	6	0,024459839	0,288635254
4	6	0,027160645	0,315795898
5	5	0,027511597	0,343307495
3	5	0,03427124	0,377578735
5	2	0,035858154	0,41343689
5	4	0,041030884	0,454467773
3	4	0,04347229	0,497940063
4	5	0,043945313	0,541885376
3	1	0,044143677	0,586029053
3	3	0,04927063	0,635299683
5	3	0,049407959	0,684707642
3	2	0,050094604	0,734802246
4	1	0,051467896	0,786270142
4	4	0,061676025	0,847946167
4	3	0,075286865	0,923233032
4	2	0,076766968	1

In Table 1 the first column contains all possible values k_1 and k_2, for which probability is $\mathrm{P}\{\eta(t_1\ t_1^*) = k_1, \eta(\ t1t) + \eta(\ t0t) = k_2\} \geq 0{,}01$. The second column of Table 1 gives the probabilities (in non-decreasing order) $P\{\eta(t_1\ t_1^*) = k_1, \eta(\ t1t) + \eta(\ t0t) = k_2\}$ for pairs of numbers (k_1, k_2) listed in the first column.

Each row of the fourth column contains the sum of the accumulated probabilities before the event is implemented $\{\eta(t_1\ t_1^*) = k_1, \eta(\ t1t) + \eta(\ t0t) = k_2\}$ inclusive where k_1 and k_2 indicated in the same line in the first column.

5.2. ILLUSTRATION OF THE USE OF EQUALITY (3)

Table 2 and Fig. 5 show the use of the relation (3) for a small sample of *n*, *n* = 16, and some values of k_1 and k_2.

Table 2. Using (3) for a small sample of length 16

k_1	k_2	P	P_c
3	6	0,01071167	0,073730469
2	5	0,011062622	0,084793091
2	1	0,011978149	0,09677124
2	0	0,012191772	0,108963013
2	4	0,012756348	0,12171936

k_1	k_2	P	P_c
2	2	0,013305664	0,135025024
2	3	0,013549805	0,148574829
4	4	0,021362305	0,169937134
3	5	0,022994995	0,192932129
6	0	0,02671814	0,219650269
5	2	0,028015137	0,247665405
3	4	0,037734985	0,285400391
4	3	0,048751831	0,334152222
3	3	0,050933838	0,38508606
3	0	0,054214478	0,439300537
3	1	0,057479858	0,496780396
3	2	0,058532715	0,55531311
5	1	0,059371948	0,614685059
4	2	0,081161499	0,695846558
5	0	0,091812134	0,787658691
4	1	0,102798462	0,890457153
4	0	0,109542847	1

Table 2 is formed of columns whose contents are similar to the contents of the Table 1 columns.

5.3. ILLUSTRATION OF THE USE OF EQUALITY (4)

Table 3 and Fig. 6 show the use of the relation (4) for a small sample n, $n = 16$, and some values k_1 and k_2.

Table 3. Using (4) for a small sample of length 16

k_1	k_2	P	P_c
5	0	0,001022339	0,003448486
1	1	0,001602173	0,005050659
6	2	0,001678467	0,006729126
6	6	0,001831055	0,008560181
2	2	0,005554199	0,01411438
5	5	0,007049561	0,02116394
6	3	0,00869751	0,02986145
1	0	0,008773804	0,038635254
6	5	0,009155273	0,047790527
3	3	0,010910034	0,058700562
4	4	0,012084961	0,070785522
5	1	0,014266968	0,08505249
6	4	0,014877319	0,09992981
4	0	0,022994995	0,122924805
2	1	0,037490845	0,160415649
5	4	0,039276123	0,199691772
2	0	0,051376343	0,251068115
5	2	0,053710938	0,304779053
3	0	0,072006226	0,376785278
5	3	0,073516846	0,450302124
3	2	0,074188232	0,524490356
4	3	0,076034546	0,600524902
4	1	0,108764648	0,709289551
3	1	0,139648438	0,848937988
4	2	0,151062012	1

Table 3 is formed of columns whose contents are similar to the contents of columns from Table 1.

5.4. ILLUSTRATION OF THE USE OF EQUALITY (5)

Table 4 shows the use of the relation (5) for a small sample n, $n = 16$, and some values k_1, k_2 and k_3.

In Table 4 in the first, second and third columns are all possible values k_1, k_2 and k_3, for which probability $\mathrm{P}\{(t_1\, t_1^*) = k_1, \eta(ttt) = k_2, \eta(tt^*t) = k_3\} \geq 0{,}01$.

Table 4. Using (5) for a small sample of length 16

k_1	k_2	k_3	P	P_c
4	0	3	0,010498047	0,303817749
6	0	4	0,010757446	0,314575195
5	0	1	0,01121521	0,325790405
5	2	3	0,011901855	0,337692261
5	0	4	0,012039185	0,349731445
3	2	2	0,012863159	0,362594604
4	0	0	0,012985229	0,375579834
5	1	4	0,014190674	0,389770508
3	4	2	0,014282227	0,404052734
3	2	0	0,014785767	0,418838501
3	3	2	0,014816284	0,433654785
5	1	2	0,015563965	0,44921875
4	3	3	0,016113281	0,465332031
3	4	1	0,016662598	0,481994629
4	1	3	0,017333984	0,499328613
3	1	0	0,018447876	0,517776489
4	2	1	0,019805908	0,537582397
4	3	2	0,019866943	0,557449341
4	2	3	0,020751953	0,578201294
3	0	0	0,021987915	0,600189209
5	1	3	0,024414063	0,624603271
3	3	1	0,025299072	0,649902344
3	0	1	0,025695801	0,675598145
3	1	1	0,029022217	0,704620361
3	2	1	0,030029297	0,734649658
5	0	2	0,033111572	0,76776123
4	1	1	0,03338623	0,801147461
5	0	3	0,033538818	0,834686279
4	2	2	0,035430908	0,870117188
4	0	2	0,040649414	0,910766602
4	1	2	0,044311523	0,955078125
4	0	1	0,044921875	1

The contents of the fourth and fifth columns are similar to the contents of the third and fourth columns of the Table 1.

5.5. RESULTS AND DISCUSSION

As a result of applying this technique for testing pseudo-random sequences for two-dimensional statistics (relations (2) – (4)), you can build a bubble diagram with which you can get the probability of the distribution of zeros and ones in a given sequence.

Consider examples of bubble diagrams for a bit sequence of small length *n*, *n* = 16.

5.6. GRAPHIC ILLUSTRATION OF THE USE OF EQUALITY (2)

Fig. 4 gives a bubble chart in which the first parameter (horizontal axis) is the value k_1, the second parameter (vertical axis) is the value k_2, and the third parameter (the bubble size) is the probability of the event occurring $\{\eta(t_1\, t_1^*) = k_1, \eta(\,t1t) + \eta(\,t0t) = k_2\,\}$, presented in percent.

After analyzing Fig. 4 it can be concluded that for the analysis of the sequence of chains of small and medium length (from 13 to 100 elements), one-dimensional statistics does not always give the correct result.

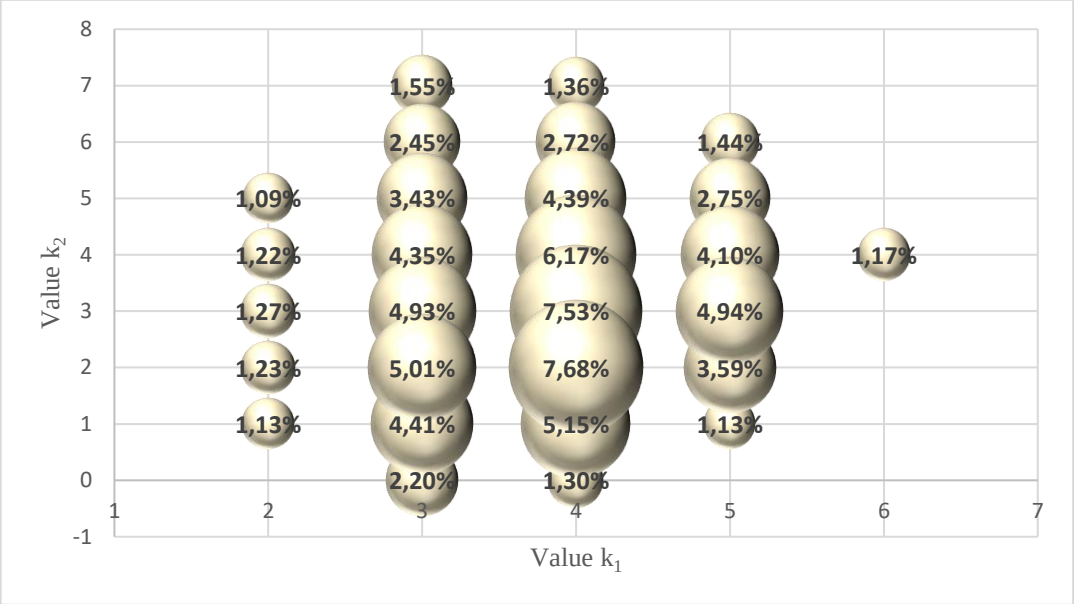

Figure 4 – Bubble chart of sequence with the length 13 for (2)

For example, if we consider the sequence where the parameter $k_1 = 4$, then we can draw a conclusion with a high degree of probability of randomness of the sequence with these characteristics, however, if we pay attention when $k_1 = 4$ and $k_2 = 0$ it can be argued that this sequence is non-random, therefore as shown in Fig. 4 we have $P\{\eta(t_1\, t_1^*) = k_1, \eta(\,t1t) + \eta(\,t0t) = k_2\} = 1{,}30\%$ that also shows the lack of use of one-dimensional statistics for the analysis of short and medium bit sequences.

An approach to testing using n-dimensional statistics allows us to rely on a deeper justification of the randomness of generated sequences.

5.7. GRAPHIC ILLUSTRATION OF THE USE OF EQUALITY (3)

Fig. 5 shows the use of the relation (3) for a small sample. n, $n = 16$, and some values k_1 and k_2.

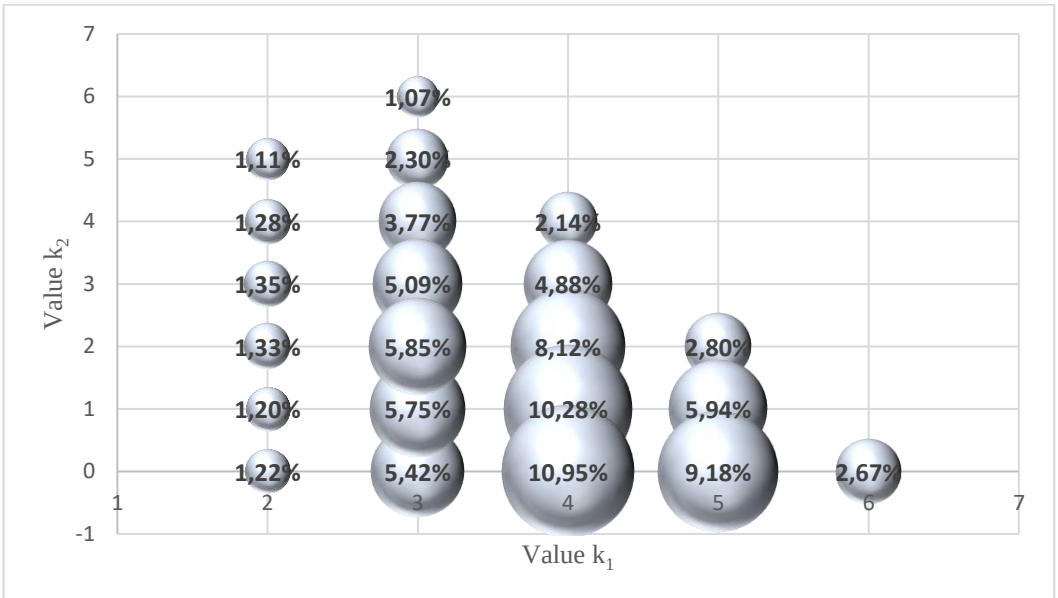

Figure 5 – Bubble chart of sequence with the length 16 for formula (3)

Fig. 5 gives a bubble chart in which the first parameter (horizontal axis) is the value k_1, the second parameter (vertical axis) is the value k_2, and the third parameter (bubble size) is the probability of the event occurring $\{\eta(t_1\ t_1^*) = k_1, \eta(\ ttt) = k_2\}$, which is represented as a percentage.

5.8. GRAPHIC ILLUSTRATION OF THE USE OF EQUALITY (4)

Fig. 6 shows the use of relation (4) for a small sample n, $n = 16$, and some values k_1 and k_2.

Fig. 6 gives a bubble chart in which the first parameter (horizontal axis) is the value k_1, the second parameter (vertical axis) is the value k_2, and the third parameter (bubble size) is the probability of the event occurring $\{\eta(t_1\ t_1^*) = k_1, \eta(\ tt^*t) = k_2\}$, which is represented as a percentage.

In this paper, the exact compatible distributions of some statistics (0, 1) -sequences of length $1 < n < \infty$ are given. For a bit sequence of small length n, n = 16, the tables containing the numerical values of the corresponding distribution are given. These tables, as well as the proposed graphic representations, can be used to test the hypothesis of the randomness of the arrangement of zeros and units.

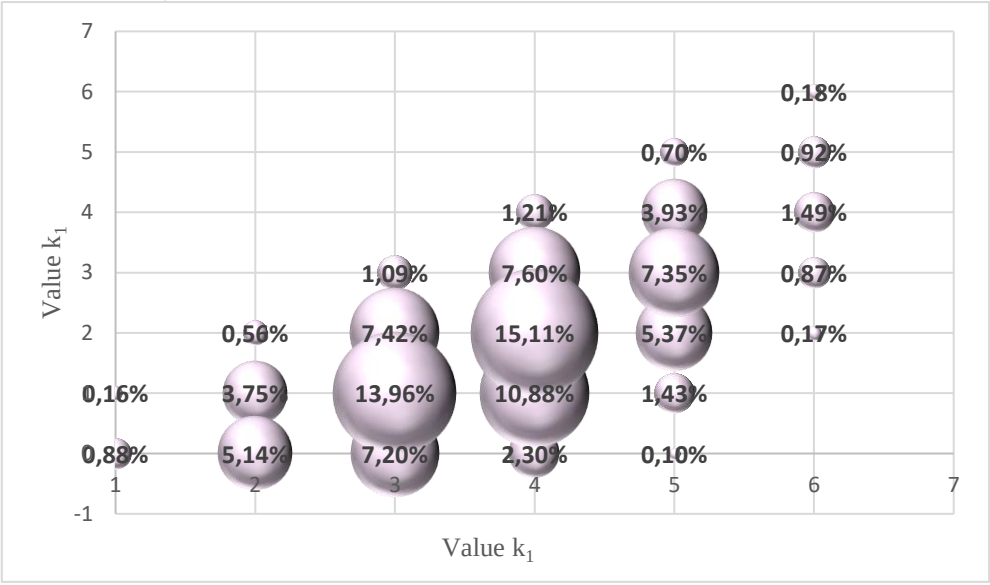

Figure 6 – Bubble chart of sequence with the length 16 for formula (4)

6. THE RESULTS OF THE COMPARISON THE NIST STATISTICAL TEST SUITE AND TEST OF PRS OF SMALL LENGTH USING MULTIDIMENSIONAL STATISTICS

Consider the well-known examples that are given in [14]. Let us analyze the submitted sequences for the corresponding tests, where:

- P is the probability of sequence randomness according to the selected criterion from the first column,
- P1 is the probability obtained using relation (2),
- P2 is the probability obtained using relation (3),
- P3 is the probability obtained using relation (4).

Table 5. The results of the comparison

Test	Input Size Recommendation	length	Sequences	P	P_1	P_2	P_3
Frequency (Monobit) Test	n>=100	10	1011010101	0,527	0,021	0,049	0,021
Frequency Test within a Block	n>=100	10	0110011010	0,801	0,097	0,212	0,129
Runs test	n>=100	10	1001101011	0,147	0,097	0,212	0,129
Binary Matrix Rank Test	n>= 38000	N=20 M = Q = 3	0101100100 1010101101	0,741	0,112	0,289	0,245
Discrete Fourier Transform (Spectral) Test	n>=1000	N=10	0001010011	0,109	0,106	0,212	0,129
Non-overlapping Template Matching Test	N >= 200	N=20, 2 blocks of length 10	1010010010 1110010110	0,344	0,098	0,176	0,105
Maurer's "Universal Statistical" Test	n>= 380000	N=20	0101101001 1101010111	0.767	0,112	0,289	0,245
Serial test	n>=100	N=10	0011011101	0,907	0,025	0,212	0,028

Test	Input Size Recommendation	length	Sequences	P	P_1	P_2	P_3
Approximate Entropy test	n>=100	N=10	0100110101	0,261	0,021	0,049	0,021
Cumulative Sums (Cusum) Test	n>=100	N=10	1011010111	0,411	0,097	0,212	0,129
Random Excursions Test	$n>10^6$	N=10	0110110101	0,502	0,003	0,049	0,003
Random Excursions Variant Test	$n>10^6$	N=10	0110110101	0,683	0,003	0,049	0,003

As can be seen from the table, the use of two-dimensional statics gives a more accurate result for short sequences. And also, according to [14], the recommended minimum sequence length n is greater than 100 bits.

7. CONCLUSION

An approach to testing the use of multidimensional statistics allows you to rely on a deeper rationale for randomized bit sequences that are being analyzed. This area is promising for scientific research. Thus, a new technique of PRS testing is proposed in the paper, and several criteria for testing bit sequence of small length are considered, which, in comparison with one-dimensional statistics, gives a more accurate result.

To implement the proposed approach, the author develops a software package for testing PRS, which will include multidimensional statistical tests.

Thus, the paper proposed a methodology for testing a memory bandwidth, and obtained a correct view of the joint distribution of the numbers of 2 chains and the numbers of 3 chains of various variants in a random bit sequence of a given small length.

To implement the proposed approach, a PRS software test package is being developed, which will include tests using multidimensional statistics, which are well recommended for testing a short length PRS. The complex is based on software products developed in C ++, Python, for analyzing PRS, as well as, the user part on a Microsoft Excel spreadsheet processor. Choosing a Microsoft Excel spreadsheet processor due to a wide segment of users, a large number of built-in mathematical and statistical functions, the possibility of programming in VBA, as well as the visibility of implementation, testing programs, there is no need to install additional programs and user training. Currently, more than 20 PRS tests have been implemented, and the test database is being updated.

An analysis of the effectiveness of pseudorandom sequence generators is an urgent issue of cybersecurity in the use of more advanced methods of encryption and information security. The available techniques show low flexibility and versatility in the means of finding hidden patterns in the data. To solve this problem, it is suggested to use algorithms based on multidimensional statistics. These algorithms combine all the advantages of statistical methods and are the only alternative for the analysis of sequences of small and medium length.

As a result of the implementation of this technique, an information system will be created that allows analyzing the PRS of a small length and choosing a quality PRS for use in a particular subject area.

8. REFERENCES

[1] A.V. Arhangelskaya, “Analysis of approaches to the definition of the term ‘randomness’,” *Proceedings of the Russian Conference “Problems of Information Security in University Education System”, MIFI-2007,* Moscow Engineering-Physical Institute, Moscow, 2007, pp. 22–23. (in Russian)

[2] S. Popereshnyak “Analysis of pseudorandom small sequences using multidimensional statistics” *Proceedings of the 2019 3rd IEEE International Conference on Advanced Information and Communication Technologies (AICT’2019),* Lviv, Ukraine, 2019, pp. 5.4.1-5.4.4.

[3] M. McLoone, J. V. McCanny, “High-performance FPGA implementation of DES using a novel method for implementing the key schedule,” *IEE Proceedings – Circuits, Devices and Systems*, vol. 150, no. 5, pp. 373-378, October 2003.

[4] F. H. Nejad, S. Sabah, A. J. Jam, “Analysis of avalanche effect on advance encryption standard by using dynamic S-Box depends on rounds keys,” *Proceedings of the 2014 International Conference on Computational Science and Technology (ICCST),* Kota Kinabalu, 2014, pp. 1-5.

[5] H. Liu, C. Jin, “Lower bounds of differential and linear active S-boxes for 3D-like structure,” *The Computer Journal*, vol. 58, no. 4, pp. 904-921, April 2015.

[6] C. U. Bhaskar, C. Rupa, “An advanced symmetric block cipher based on chaotic systems,” *Proceedings of the 2017 Innovations*

in Power and Advanced Computing Technologies (i-PACT), Vellore, 2017, pp. 1-4.

[7] N. Ferguson, B. Schneier, *Practical Cryptography*, John Wiley & Sons, 2003, 432 p.

[8] A.J. Menezes, P.C. van Oorschot, S.A. Vanstone, *Handbook of Applied Cryptography*, CRC Press, 1997, 794 p.

[9] B. N. Tran, T. D. Nguyen and T. D. Tran, "A new S-box structure to increase complexity of algebraic expression for block cipher cryptosystems," *Proceedings of the 2009 International Conference on Computer Technology and Development*, Kota Kinabalu, 2009, pp. 212-216.

[10] P. Busireddygari, S. Kak, "Pseudorandom tableau sequences," *Proceedings of the IEEE 51st Asilomar Conference on Signals, Systems, and Computers*, 2017, pp. 1733-1736.

[11] S. Gurugopinath, B. Samudhyatha, "Multi-dimensional AndersonDarling statistic based goodness-of-fit test for spectrum sensing," *Proceedings of the IEEE Seventh International Workshop on Signal Design and its Applications in Communications (IWSDA)*, Bengaluru, India, 2015, pp. 165-169.

[12] H. Wang, E.-H. Yang, Z. Zhao, W. Zhang, "Spectrum sensing in cognitive radio using goodness of fit testing," *IEEE Transactions on Wireless Communications*, vol. 8, issue 11, pp. 5427-5430, 2009.

[13] D. Teguig, V. Le Nir, B. Scheers, "Spectrum sensing method based on goodness of fit test using chi-square distribution," *Electronics Letters*, vol. 50, issue 9, pp. 713-715, 2014.

[14] Special Publication 800-22, A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications. [Online]. Available at: http://csrc.nist.gov.

[15] The eSTREAM Project, 2004, [Online]. Available at: http://www.ecrypt.eu.org.

[16] ISO/IEC 18033-4:2011. Information technology – Security techniques – Encryption algorithms – Part 4: Stream ciphers, 2012.

[17] D. D. Ismoyo, R. W. Wardhani, "Block cipher and stream cipher algorithm performance comparison in a personal VPN gateway," *Proceedings of the 2016 International Seminar on Application for Technology of Information and Communication (ISemantic)*, Semarang, 2016, pp. 207-210.

[18] D. Moody, "Post-quantum cryptography: NIST's plan for the future," *Proceedings of the Seventh International Conference on Post Quantum Cryptography*, Japan, 2016. [Online]. Available at: https://pqcrypto2016.jp.

[19] The Marsaglia, "Random Number CDROM including the Diehard Battery of Tests of Randomness," [Online]. Available at: http://stat.fsu.edu/pub/diehard.

[20] eSTREAM Optimized Code HOWTO, 2005. [Online]. Available at: http://www.ecrypt.eu.org.

[21] M. Robshaw, O. Billet, "New stream cipher designs: The eSTREAM," *Finalists*, Berlin, 2008.

[22] A. A. Zadeh, H. M. Heys, "Application of simple power analysis to stream ciphers constructed using feedback shift registers," *The Computer Journal*, vol. 58, no. 4, pp. 961-972, April 2015.

[23] C. Carlet et al., "Analysis of the algebraic side channel attack," *Journal of Cryptographic Engineering*, vol. 1, no. 2, pp. 45-62, 2012.

[24] A. R. Kazmi, M. Afzal, M. F. Amjad, A. Rashdi, "Combining algebraic and side channel attacks on stream ciphers," *Proceedings of the 2017 International Conference on Communication Technologies (ComTech)*, Rawalpindi, 2017, pp. 138-142.

[25] D. P. Upadhyay, P. Sharma, S. Valiveti, "Randomness analysis of A5/1 Stream Cipher for secure mobile communication," *International Journal of Computer Science & Communication*, vol. 3, pp. 95-100, 2014.

[26] D. Upadhyay, T. Shah, P. Sharma, "Cryptanalysis of hardware based stream ciphers and implementation of GSM stream cipher to propose a novel approach for designing n-bit LFSR stream cipher," *Proceedings of the 2015 19th International Symposium on VLSI Design and Test*, Ahmedabad, 2015, pp. 1-6.

[27] P. Pillai, S. Pote, "Physical layer security using stream cipher for LTE," *Proceedings of the 2015 IEEE Bombay Section Symposium (IBSS)*, Mumbai, 2015, pp. 1-5.

[28] C. Cassisi, P. Montalto, M.A. Aliotta, A. Pulvirenti, "Similarity measures and dimensionality reduction techniques for time series data mining," *Advances in Data Mining Knowledge Discovery and Applications*, Chapter 3, IntechOpen, London, 2012, pp. 71-96.

[29] D. Berndt, J. Clifford, "Using dynamic time warping to find patterns in time series," *Workshop on KDD*, vol. 10, no. 16, Seattle, USA, July 31 – August 01, 1994, pp. 359-370.

[30] 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*, vol. 76, pp. 103-116, 2008.

[31] I.P. Gaydyshev, *Data analysis software*, AtteStat. User's manual. Version 13, 2012, 525 p. (in Russian)

[32] S. Popereshnyak, G. P. Dimitrov, "The testing of pseudorandom sequences using multidimensional statistics," [Online]. Available at: ceur-ws.org/Vol-2533/paper14.pdf.

[33] V. Masol, S. Popereshnyak "Statistical analysis of local sections of bits sequences," *Journal of Automation and Information Sciences*, vol. 51, issue 10, pp. 31-45, 2019. DOI: 10.1615/JAutomatInfScien.v51.i10.30.

[34] S. Popereshnyak, "The technique for testing short sequences as a component of cryptography on the Internet of Things," [Online]. Available at: http://ceur-ws.org/Vol-2516/paper11.pdf.

***Svitlana Popereshnyak**, Candidate of Physical and Mathematical Sciences, an Associate Professor at the Department of Software Systems and Technologies at Taras Shevchenko National University of Kyiv.*

Areas of scientific interests: software engineering, probability theory and mathematical statistics, applied cryptology, methods of information security in computer systems.