# The Testing of Pseudorandom Sequences using Multidimensional Statistics

Svitlana Popereshnyak <sup>1[0000-0002-0531-9809]</sup> and Georgi P. Dimitrov <sup>2[0000-0001-5064-3168]</sup>

<sup>1</sup> Taras Shevchenko National University of Kyiv, 24, Bohdana Havrylyshyna str., Kyiv, 04116, Ukraine

spopereshnyak@gmail.com

<sup>2</sup> University of Library Studies and Information Technologies, 119, Tsarigradsko Shose,
Sofia, Bulgaria

geo.p.dimitrov@gmail.com

**Abstract.** The available approaches to testing pseudorandom sequences 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. The paper proposed a new approach for testing pseudorandom sequences, obtained an explicit form of the joint distribution of numbers of 2-chains and numbers of 3-chains of various options random bit sequence of a given small length. Examples, tables, diagrams that can be used to test for randomness of the location of zeros and ones in the bit section are presented. In future as a result 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.

**Keywords:** Algorithms, multidimensional Statistics, Random Sequence, *s*-chains, Cryptography, Pseudorandom Sequence, Statistical Testing.

# 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: modeling, cryptography and information security, decision making in automated expert systems, optimization of functional dependencies, fun and games.

There are various approaches to the formal definition of the term "randomness" based on the concepts of computability and algorithmic complexity [1-2].

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.

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0) 2019 DCSMart Workshop.

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 [3-6].

#### 2 Problem Statement

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

- checked out only one of the probable ones properties that are characterize PRS;
- not fix family alternatives;
- do not have theoretical ones ratings power.
- do not give a correct an estimate of chance sequences provided a little sample.

Problems small and large samples refer to the main problems that arise in practical application methods analysis data. Let's be use the next classification samples by number [2], based on requirements presented in the program criteria:

- very small sample from 5 to 12,
- small sample from 13 to 40,
- medium sample from 41 to 100,
- large sample 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. For a 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.

# 3 Joint Distribution of number of 2-chains and number of 3chains of a provided type in binary sequence

Consider a sequence of random variables

$$\gamma_1, \gamma_2, \dots, \gamma_n,$$
 (1)

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

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;  $P\{\gamma_i = 1\} = p$ ,  $P\{\gamma_i = 0\} = q$ , p + q = 1,  $i = 1, 2, \ldots, n$  and  $k_1, k_2, k_3, t$ , — integer numbers such that  $k_1 \ge 0, k_2 \ge 0, k_3 \ge 0$ ,  $m_1 + m_0 = n \ge 3$ ,  $t \in \{0, 1\}$ ,  $t^* = 1 - t$ . Then

 $+\chi(m_t-k_1-1,k_2,m_{t^*})\},$ 

is the symbol  $\sum$  denotes addition over all non-negative integers  $\delta_t$  and  $\delta_{t^*}$  such that  $a_1=k_2-m_{t^*}+2(m_t-k_1-1), \quad a_2=k_2-m_{t^*}+2(m_t-k_1+1), \quad a_1=k_2-m_{t^*}+2(m_t-k_1).$ 

# 4 Experiment

As a result of applying this technique for testing pseudo-random sequences for two-dimensional statistics, you can build tables (relations (2) - (5)) and bubble diagrams (relations (3) - (5)) with which you can get the probability of the distribution of zeros and ones in a given sequences.

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 and bubble diagrams for a bit-sequence of small length. For example, let the length of the bit sequence n, n = 32 for relations (3) - (5) and n = 24 for relations (2).

#### 4.1 Illustration of the Use of Equality (2)

In Table 1 and in Fig. 1 shows the use of the relation (2) for a small sample n, n = 32, and some values  $k_1$  and  $k_2$ .

**Table 1.** Using relation (3) for a small sample of length 32

$k_1$	$k_2$	P	$P_c$	$k_1$	$k_2$	P	$P_c$
4	5	0,0102	0,44366	9	4	0,01595	0,67931
6	1	0,01037	0,45403	6	2	0,01596	0,69527
12	1	0,0108	0,46483	10	1	0,01623	0,7115
5	2	0,01106	0,4759	8	1	0,01642	0,72791
9	5	0,01121	0,48711	6	5	0,01655	0,74446
11	3	0,01157	0,49868	7	5	0,01655	0,76102
5	6	0,01187	0,51055	9	1	0,01721	0,77823
10	4	0,01189	0,52244	10	2	0,0181	0,79633
7	6	0,01203	0,53447	6	4	0,01898	0,81531
6	6	0,01289	0,54736	6	3	0,01901	0,83432
11	1	0,01387	0,56123	8	4	0,01915	0,85346
7	1	0,01393	0,57516	7	2	0,01981	0,87328
5	3	0,01417	0,58933	9	3	0,01985	0,89313
5	5	0,0142	0,60353	7	4	0,02039	0,91351
11	2	0,01421	0,61774	9	2	0,02085	0,93437
8	5	0,01449	0,63222	8	2	0,02156	0,95593
5	4	0,01519	0,64741	7	3	0,02192	0,97785
10	3	0,01595	0,66336	8	3	0,02215	1

In Table 1 the first column contains all possible values  $k_1$  and  $k_2$ , for which probability is  $P\{\eta(t\,t)=k_1,\eta(\,t^*t^*t^*)=k_2\}\geq 0.01$ . The second column of Table 1 gives the probabilities (in non-decreasing order)  $P\{\eta(t\,t)=k_1,\eta(\,t^*t^*t^*)=k_2\}$  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  $\{\eta(t\,t)=k_1,\eta(\,t^*t^*t^*)=k_2\}$  inclusive where  $k_1$  and  $k_2$  indicated in the same line in the first column.

# 4.2 Illustration of the Use of Equality (4)

In Table 2 and in Fig. 2. shows the use of the relation (4) for a small sample of n, n = 32, and some values of  $k_1$  and  $k_2$ .

Table 2. Using relation (4) for a small sample of length 32

$k_1$	$k_2$	Р	$P_{c}$	$k_1$	$k_2$	P	$P_{c}$
$\frac{-\kappa_1}{12}$	1	0,010309	0,25025	5	6	0,019461	0,524834
			,			,	
4	7	0,010346	0,260596	11	2	0,020707	0,545541
13	2	0,010566	0,271162	8	2	0,020939	0,56648
10	1	0,010906	0,282067	6	3	0,022517	0,588997
11	1	0,011296	0,293363	5	4	0,023782	0,61278
3	6	0,011426	0,304789	10	2	0,024014	0,636794
7	6	0,01148	0,316269	9	2	0,024221	0,661015
9	5	0,011732	0,328001	9	4	0,025878	0,686893
12	3	0,011875	0,339876	7	5	0,026396	0,713288
5	3	0,013051	0,352927	10	3	0,027086	0,740375
4	4	0,013083	0,36601	5	5	0,027095	0,76747
7	2	0,015224	0,381234	6	5	0,029893	0,797363
12	2	0,015705	0,396939	7	3	0,030948	0,828311
6	6	0,016693	0,413631	6	4	0,033093	0,861404
10	4	0,017033	0,430665	9	3	0,033247	0,894651
4	6	0,017494	0,448159	8	4	0,033621	0,928272
4	5	0,018859	0,467018	8	3	0,034964	0,963236
11	3	0,019157	0,486174	7	4	0,036764	1
8	5	0,019199	0,505373				

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

#### 4.3 Illustration of the Use of Equality (5)

In Table 3 and in Fig. 3 shows the use of the relation (5) for a small sample n, n = 32, and some values  $k_1$  and  $k_2$ .

**Table 3.** Using relation (5) for a small sample of length 32

$k_1$	$k_2$	P	$P_c$	$k_1$	$k_2$	Р	$P_c$
6	11	0,01018	0,35129	6	10	0,01737	0,61156
4	12	0,01025	0,36154	10	6	0,019	0,63056
6	6	0,01028	0,37182	5	10	0,01963	0,65019
12	4	0,01165	0,38347	7	9	0,01986	0,67005
7	10	0,01178	0,39525	7	6	0,02017	0,69022
11	6	0,01179	0,40704	6	7	0,02026	0,71048
9	4	0,01209	0,41913	10	5	0,02064	0,73112
4	9	0,01229	0,43143	9	7	0,02083	0,75195
8	9	0,01285	0,44428	8	8	0,0211	0,77305
10	7	0,0129	0,45718	9	5	0,02156	0,79461
9	8	0,01325	0,47043	5	9	0,02159	0,8162
5	11	0,014	0,48444	6	9	0,0242	0,8404
4	11	0,01416	0,4986	9	6	0,02513	0,86552
11	4	0,01481	0,51341	6	8	0,02612	0,89165
10	4	0,01521	0,52862	8	6	0,02619	0,91783
4	10	0,01543	0,54406	7	8	0,02698	0,94481
11	5	0,01578	0,55984	8	7	0,02735	0,97217
8	5	0,01706	0,57691	7	7	0,02783	1
5	8	0,01729	0,5942				
		-		·	·		

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

# 4.4 Illustration of the Use of Equality (2)

In Table 4 shows the use of the relation (2) for a small sample n, n=24, and some values  $k_1$ ,  $k_2$  and  $k_3$ .

**Table 4.** Using relation (2) for a small sample of length 24

$k_1$	$k_2$	$k_3$	P	$P_c$
5	1	3	0,009096	0,851162
4	4	3	0,009398	0,86056
5	1	4	0,009748	0,870309
8	1	2	0,009901	0,88021
7	1	3	0,009946	0,890155
4	3	3	0,009999	0,900154
6	3	2	0,010374	0,910529
7	1	2	0,010382	0,920911
4	2	4	0,010422	0,931332
6	2	2	0,010553	0,941885
7	2	2	0,011017	0,952902
5	3	3	0,011284	0,964186
6	2	3	0,011495	0,975681
6	1	3	0,011903	0,987584
5	2	3	0,012416	1

In Table 4 in the first, second and third columns are all possible values  $k_1$ ,  $k_2$  and  $k_3$ , for which probability  $P\{\eta(t\,t)=k_1,\eta(\,t^*t^*t^*)=k_2,\eta(\,t^*t\,t^*)=k_3\,\}\geq 0{,}009$ , and the contents of the fourth and fifth columns are similar to the contents of the third and fourth columns of the Table 1.

# 5 Results and Discussion

As a result of applying this technique for testing pseudo-random sequences for twodimensional statistics (relations (3) - (5)), 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 = 32.

# 5.1 Graphic Illustration of the Use of Equality (3)

Fig. 1 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  $\{\eta(t\,t)=k_1,\eta(\,t^*t^*t^*)=k_2\}$ , presented in percent.

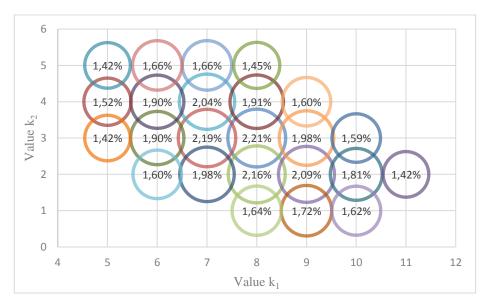


Fig. 1. Bubble chart of sequence with the length 32 for (3)

After analyzing Fig. 1 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 do not always give the correct result. For example, if we consider the sequence where the parameter  $k_1 = 8$ , then we can draw a conclusion with a degree of probability about 10% of randomness of the sequence with these characteristics, however, if we pay attention when  $k_1 = 8$  and  $k_2 = 5$  it can be argued that this sequence is non-

random, therefore as shown in Fig. 1 we have  $P\{\eta(t\,t)=k_1,\eta(\,t^*t^*t^*)=k_2\}=1,45\%$ . What also shows the lack of use of one-dimensional statistics for the analysis of small 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.

# 5.2 Graphic Illustration of the Use of Equality (4)

In Fig. 2 shows the use of the relation (4) for a small sample  $n,\,n=32$  , and some values  $k_1$  and  $k_2$ .



Fig. 2. Bubble chart of sequence with the length 32 for formula (4).

Fig. 2 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  $\{\eta(t\,t)=k_1,\eta(\,t^*t\,t^*)=k_2\,\}$ , which is represented as a percentage.

# 5.3 Graphic Illustration of the Use of Equality (5)

In Fig. 3 shows the use of relation (4) for a small sample n, n = 32, and some values  $k_1$  and  $k_2$ .

Fig. 3 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  $\{\eta(t\,t)=k_1,\eta(\,t^*t^*t^*)+\eta(\,t^*t\,t^*)=k_2\,\}$ , which is represented as a percentage.

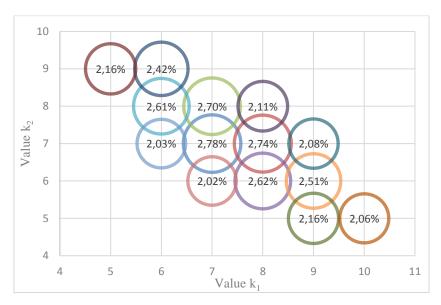


Fig. 3. Bubble chart of sequence with the length 32 for formula (5).

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 = 32, 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.

# 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 [7, 8]. 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.
- P<sub>1</sub> is the probability obtained using relation (2),
- P<sub>2</sub> is the probability obtained using relation (3),
- P<sub>3</sub> is this is the probability obtained using relation (4),
- $P_4$  is this is the probability obtained using relation (5).

**Table 5.** The results of the comparison

Test	Input Size Recom- mendation, n more than	length	Sequences	P	P <sub>1</sub>	P <sub>2</sub>	P <sub>3</sub>	P <sub>4</sub>
Frequency (Monobit) Test	100	10	1011010101	0,527	0,007	0,027	0,007	0,057
Frequency Test within a Block	100	10	0110011010	0,801	0,01	0,075	0,102	0,01
Runs test	100	10	1001101011	0,147	0,052	0,075	0,087	0,09
Binary Matrix Rank Test Discrete	38000	N=20 M = Q = 3	01011001001 010101101	0,741	0,004	0,008	0,014	0,017
Fourier Transform (Spectral) Test	1000	N=10	0001010011	0,109	0,063	0,109	0,084	0,092
Non- overlapping Template Matching Test	200	N=20, 2 blocks of length 10	10100100101 110010110	0,344	0,01	0,026	0,051	0,025
Maurer's "Universal Statistical" Test	380000	N=20	01011010011 101010111	0.767	0,001	0,03	0,009	0,023
Serial test	100	N=10	0011011101	0,907	0,029	0,064	0,087	0,088
Approximate Entropy test Cumulative	100	N=10	0100110101	0,261	0,052	0,075	0,087	0,09
Sums (Cusum) Test	100	N=10	1011010111	0,411	0,02	0,031	0,043	0,057
Random Excursions Test	10 <sup>6</sup>	N=10	0110110101	0,502	0,02	0,027	0,043	0,031
Random Excursions Variant Test	10 <sup>6</sup>	N=10	0110110101	0,683	0,02	0,027	0,043	0,031

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 [8], the recommended minimum sequence length n is greater than 100 bits.

# 7 Conclusions

The available approaches to testing pseudorandom sequences 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.

The approach to testing using multidimensional statistics allows you to rely on a deeper justification of the randomness of the generated sequences. This area is promising for scientific research.

The paper proposed a methodology for testing a sequence 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.

These algorithms and scheme of work for verification statistical tests of randomness sequences (proposed in chapter II) combine all the advantages of statistical methods and are the only alternative for the analysis of sequences of small and medium 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 small length PRS. As a result of the implementation of this technique, an information system will be created that will allow analyzing the PRS of a small length and choosing a quality PRS for use in a particular subject area.

#### References

- Masol V., Popereshnyak S. Statistical analysis of local plots of bits sequences. Problemy upravleniya i informatiki, 5, 92-105 (2019).
- Popereshnyak S. Analysis of pseudorandom small sequences using multidimensional statistics. In: The 3<sup>rd</sup> IEEE International Conference on Advanced Information and Communication Technologies (AICT), pp. 5.4.1-5.4.4. IEEE Press, Ukraine (2019)
- 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)
- 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)
- 5. Busireddygari P.; Kak S. Pseudorandom tableau sequences, In: 51st Asilomar Conference on Signals, Systems, and Computers, pp. 1733 1736. IEEE Press (2017)
- 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).
- 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
- 8. Special Publication 800-22. A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications. http://csrc. nist.gov