Журнал Белорусского государственного университета. Математика. Информатика. 2026;1:22–35
Journal of the Belarusian State University. Mathematics and Informatics. 2026;1:22–35

УДК 519.233.3 + 519.719.2

НАДЕЖНОСТЬ ДВУХУРОВНЕВОГО ПОДХОДА К ТЕСТИРОВАНИЮ В НАБОРЕ ТЕСТОВ NIST SP 800-22

А. А. СЕРОВ[1)]

[1)]*Математический институт им. В. А. Стеклова РАН, ул. Губкина, 8, 119333, г. Москва, Россия*

Аннотация. Рассматривается предложенный в наборе тестов NIST SP 800-22 двухуровневый подход к тестированию генераторов случайных чисел, который состоит в подсчете количества последовательностей, прошедших основной тест, и проверке распределения *p*-значений с помощью хи-квадрат-теста. Данный подход может повысить надежность статистического теста, однако он чувствителен к погрешностям аппроксимации, возникающим при вычислении *p*-значений. Показано, что для последовательностей, порождаемых с помощью симметричного алгоритма блочного шифрования (AES), двухуровневый подход к тестированию не является надежным. Систематическая погрешность при вычислении *p*-значений зависит только от погрешности аппроксимации точного распределения статистики эквивалентным ему теоретическим распределением и от числа бит *n* в тестируемых последовательностях. Для надежного двухуровневого тестирования такая погрешность не должна превосходить $\frac{\sigma}{N} = \frac{1}{k}\sqrt{\frac{k-1}{N}}$, где $\sigma = \sqrt{\frac{1}{k}\left(1-\frac{1}{k}\right)N}$ – стандартное отклонение от среднего числа частиц в ячейке в равновероятной схеме размещения N частиц по k ячейкам. Подобные эвристические соображения и проведенные эксперименты позволяют предположить, что, например, при $n = 2^{20}$ в двухуровневом частотном тесте (*Frequency test*) из набора тестов NIST SP 800-22 количество тестируемых последовательностей N не должно превышать 26 000. Для полного исключения систематической погрешности, возникающей при проведении частотного теста и определении номера ячейки из k возможных ячеек (непересекающихся подынтервалов отрезка $[0,1]$), которому принадлежит *p*-значение, предложены двусторонние оценки квантилей биномиального закона.

Ключевые слова: случайная последовательность; псевдослучайная последовательность; статистическое тестирование; надежность статистического теста; биномиальное распределение; двусторонняя оценка.

RELIABILITY OF TWO-LEVEL TESTING APPROACH IN THE NIST SP 800-22 TEST SUITE

A. A. SEROV[a]

[a]*Steklov Mathematical Institute, Russian Academy of Sciences, 8 Gubkina Street, Moscow 119333, Russia*

Abstract. The two-level approach for testing random number generators involving the well-known NIST SP 800-22 test suite, i. e. counting the sequences passing a basic test and checking the *p*-values distribution with a chi-square test, was considered. Such an approach may increase the reliability of the test. However, it is sensitive to the approximation error introduced by the computing of *p*-values. In this paper it is shown that for AES-based sequences two-level testing approach is not reliable too. Systematic error in the computing of the *p*-values is dependent only on the ac-

Образец цитирования:
Серов АА. Надежность двухуровневого подхода к тестированию в наборе тестов NIST SP 800-22. *Журнал Белорусского государственного университета. Математика. Информатика.* 2026;1:22–35 (на англ.).
EDN: BPTBQY

For citation:
Serov AA. Reliability of two-level testing approach in the NIST SP 800-22 test suite. *Journal of the Belarusian State University. Mathematics and Informatics.* 2026;1:22–35.
EDN: BPTBQY

Автор:
Александр Александрович Серов – кандидат физико-математических наук; научный сотрудник отдела дискретной математики.

Author:
Aleksandr A. Serov, PhD (physics and mathematics); researcher at the department of discrete mathematics.
serov@mi-ras.ru
https://orcid.org/0000-0001-7465-1701

curacy of approximation of the exact distribution of statistic by its theoretical counterpart and the number of bits in the analysed sequences n. For a reliable second-level test, this error should be smaller, or at least, approximately equal to $\frac{\sigma}{N} = \frac{1}{k}\sqrt{\frac{k-1}{N}}$, where $\sigma = \sqrt{\frac{1}{k}\left(1-\frac{1}{k}\right)N}$ is the standard deviation from the mean number of particles in a bin for equiprobable scheme of allocation N particles in k bins. Such heuristic assumptions and carried out experiments suggest that for example in the second-level test of the Frequency test of NIST SP 800-22 test suite with $n = 2^{20}$ the number of tested sequences N should not exceed 26 000. To completely eliminate the systematic error appearing in the Frequency test when determining the number of bin from k bins (disjoint subintervals of the interval $[0,1]$) to which the p-value belongs the two-sided estimates of the quantiles of the binomial law are proposed.

Keywords: random sequence; pseudorandom sequence; statistical testing; reliability of statistical test; binomial distribution; two-sided estimate.

Introduction

Random sequences are used in a large variety of areas, such as quantum mechanics, game theory, statistics, cryptography and so on. These sequences may be generated either by physical sources or deterministic algorithms. Random number generators (RNGs) represent a fundamental component in many applications, they are essential for cryptographic systems (see, for example, [1]). The security of many cryptographic schemes and protocols is based on the perfect randomness of RNG outputs. RNGs are classified into two types – pseudo (deterministic) and true (non-deterministic) RNGs (further – PRNGs and TRNGs, respectively). In general, TRNGs based on some random physical phenomenona, may be used directly as random bit sources or generate seeds for PRNGs, and PRNGs extend the seeds to produce deterministic long sequences.

For any type of RNG, statistical hypothesis tests have been widely employed to assess the quality of the RNG, which evaluate whether the output sequences conform with the given null hypothesis (e. g., the elements of the sequence independent and uniformly distributed) or not. In addition, statistical randomness tests are also used to evaluate the outputs of other cryptographic primitives such as block ciphers and hash functions, to preliminarily validate the indistinguishability of their outputs from a equiprobable random mapping. The quality checking of binary sequences usually is based on some well-known batteries of tests, each of which is composed of a serial of tests, include Diehard[1] proposed by G. Marsaglia, SP 800-22 [2] standardised by National Institute of Standards and Technology (further – NIST SP 800-22) or a software library TestU01 [3].

Note that a test for randomness may be interpreted only in a probabilistic way. Looking at a sequence of all zeros, we say that the sequence is not random at all, though for an ideal RNG this sequence has the same probability of any other sequence of the same length; on the contrary, a RNG that is not able to generate this sequence is not ideal.

The main problem of the statistical testing is the interpretation of the results. Roughly speaking, while a failed test is a serious indicator for the weakness of a RNG, a passed test does not provide a direct positive proof for the quality of a RNG. From a mathematical point of view a test may be considered as a function of a sequence of n elements (e. g., a sequence of n bits) with output value in the interval $[0,1]$, called a p-value. In null-hypothesis significance testing, the p-value is the probability of obtaining test results at least as extreme as the result actually observed, under the assumption that the null hypothesis is correct.

In general case, if the null hypothesis is true, the p-value based on a continuous test statistic has a uniform distribution over the interval $[0,1]$, regardless of the sample size of the experiment. In contrast, the distribution of the p-value under the alternative hypothesis is a function of both sample size and the true value or range of true values of the tested parameter of the true distribution. In the discrete case, then the test statistic distribution is not continuous and the null hypothesis is true, i. e. all elements of the input sequence are independent and drawn according to the uniform distribution, the cumulative distribution function (CDF) $F_p(x)$ of p-value coincides with the value of its argument in a discrete set of points from the interval $[0,1]$. If the elements of the generated sequence is non-independent or not distributed according to the uniform distribution, then the CDF of p-value is not known as a rule.

NIST SP 800-22 test suite

In this paper the most commonly used statistical test suite, the NIST SP 800-22 test suite [2], is considered. It is well known that the NIST statistical test suite was used for the evaluation of AES candidate algorithms. This

[1] *Marsaglia G.* Diehard: a battery of tests of randomness [Electronic resource]. URL: http://stat.fsu.edu/geo/diehard.html (date of access: 21.06.2025).

statistical test suite is build for analysing the randomness properties of sequences and generators, is composed of 15 tests, and provides comprehensive evaluation for different randomness aspects of assessed sequences. All NIST SP 800-22 tests are listed in table 1.

Table 1

All NIST SP 800-22 tests with brief descriptions

Number	Test name	Used statistics
1	Frequency test	Normalised modulus of the difference between frequencies of 1 and 0
2	Block frequency test	Chi-square statistic of 1 frequencies in adjacent non-intersecting 128-bit blocks
3	Cumulative sums test	Maximal deviation from 0 for partial sums of ±1 walk constructed by the binary sequence
4	Runs test	The total number of 1-runs and 0-runs
5	Longest run test	Pearson statistic of maximal lengths of 1-runs frequencies in adjacent non-intersecting 10^4-bit blocks
6	Binary matrix rank test	Pearson statistic of binary 32 × 32-matrices ranks of frequencies formed from adjacent non-intersecting 1024-bit blocks of the sequence
7	Discrete Fourier transform test	Normalised difference between the number of discrete Fourier transform coefficients of n-bit sequence, exceeding $\sqrt{n \ln 20}$ in absolute value, and $\frac{0.95n}{2}$
8	Overlapping template matching test	Pearson statistic of 9-bit 1-series frequencies in 1032-bit adjacent blocks with overlappings
9	Maurer's universal statistical test	Sum of base 2 logarithms of distances between equal non-intersecting 7-bit blocks
10	Approximate entropy test	Difference of logarithmic frequencies statistics of 10- and 11-bit segments with overlappings, the reference distribution is the χ^2 distribution
11	Serial test	Differences in Pearson statistics of frequencies of all 14-, 15- and 16-bit segments with overlappings
12	Linear complexity test	Pearson statistic of shortest LSR lengths frequencies generating 500-bit segments of the sequence
13	Non-overlapping template matching test	Chi-square statistics of aperiodic fixed segments frequencies in 8 adjacent non-overlapping blocks
14	Random excursions test	Pearson statistics of random walk cycles frequencies with fixed numbers visiting of –4, –3, …, 4 states
15	Random excursions variant test	The total number of times that the given 18 states from –9 to 9 is visited during the entire random walk

The main reason for using the NIST SP 800-22 test suite, from an engineering point of view, is that it has several appealing properties. First, the suite is composed of several different tests, each of them is applied to the same input sequence of n bits (for many of the tests the assumption has been made that the sequence length n is a value from 10^3 to 10^7), searching for a specific statistical feature and expressing it as numerical quantity of p-value[2]. Second, the suite is composed by a number of well-known tests and, for all of them, an exhaustive mathematical treatment is available. The source code of all the tests in the suite is public available.

The purpose of this paper is to consider the testing strategy proposed in section 4 of the NIST publication [2] and discuss under which assumptions this strategy increases the reliability and when, on the other hand, produces incorrect results, i. e. the empirical significance level α does not correspond to the theoretical one. In that context, a reliable test should be understood as a test such that the probability of a false positive (type I error) is agreed with the expected one.

In the beginning of the testing process, the whole bit sequence is divided into N blocks of the length n bits, then a test statistic value is computed for each block. All 15 tests may be divided into two types (according to the test statistic distribution) – the binomial-based (i. e. two-sided) and the chi-square-based (i. e. one-sided) tests. The Frequency test, the Runs test, the Discrete Fourier transform test, Maurer's universal statistical test, the

[2]Actually, the Non-overlapping template matching test, the Random excursions test, the Random excursions variant test, the Serial test and the Cumulative sums test generate 148, 8, 18, 2 and 2 p-values, respectively, other tests – 1 p-value each; however, it is very common considering only 1 of p-value from the set of p-values related to the test.

Random excursions variant test and the Cumulative sums test belong to the binomial-based tests, and the others are chi-square-based tests.

For each value of test statistic the p-value is computed according to the theoretic test statistic distribution[3]. A test is considered to be passed when the computed p-value is larger than the significance level α. According to the computed p-values for N blocks, each test performs two-level test – the first-level and the second-level tests. The first-level test focuses on the passing ratio of the N p-values, and the second-level test further focuses on the uniformity of the N p-values to assess whether the test statistic values follow the expected distribution, i. e. the chi-square distribution. The two-level testing approach was found to increase the testing capability [4].

Related works

There are many articles on the NIST SP 800-22 test suite. Among them S.-J. Kim, K. Umeno and A. Hasegawa in work [5] have found that the test setting of the Discrete Fourier transform test and Lempel – Ziv test from an earlier version of NIST SP 800-22 suite are wrong, they give four corrections of mistakes in the test settings, K. Hamano and T. Kaneko in papers [6; 7] derived the distribution function of the discrete Fourier transform spectrum, changed the threshold value from the default value of $\sqrt{3n}$ to the value of $\sqrt{\left(\ln\frac{1}{0.05}\right)n}$, where n is the length of a random number sequence, and corrected the occurrence probabilities π_i of the Overlapping template matching test included. F. Sulak, A. Doğanaksoy, B. Ege and O. Koçak in article [8] found that the p-values for short sequences (less than 512 bits) follow a specific discrete distribution, rather than the assumed uniform distribution for long sequences, calculated the probabilities of subintervals for some NIST test and proposed an alternative approach to evaluate test results for short sequences. F. Pareschi, R. Rovatti and G. Setti in work [4] investigated the reliability of the second-level tests, and analysed the sensitivity to the approximation errors introduced by the computation of p-values. Since the sequence length is finite in practice and thus the set of possible statistic values is discrete, F. Pareschi et al. in paper [9] provided the more closer distributions to the actual ones of p-values for the Frequency test, the Runs test, the Discrete Fourier transform test.

Recently, Yu. S. Kharin and A. M. Zubkov in article [10] proposed a model of a complex null hypothesis allowing possible deviations for the distribution of output sequences of generators from the equiprobable distribution, and developed an approach to the construction of the criterion for testing such hypothesis. The results obtained in work [10] may be applied in practice in statistical testing of physical random sequence generators.

First-level testing approach of the NIST SP 800-22

Denote by $\mathcal{H}_0$ the null hypothesis on the uniformity and independence of binary sequence elements, i. e. that elements of binary sequence are independent and take values 0 and 1 with probabilities $\frac{1}{2}$. In other words, hypothesis $\mathcal{H}_0$ is that the RNG under test produces sequences with the above properties, such a generator will be called an ideal RNG further. In the statistical hypothesis testing, a test statistic is chosen and used to determine whether $\mathcal{H}_0$ should be accepted or rejected. Under the null hypothesis $\mathcal{H}_0$, the theoretical reference distribution of the statistic is determined by the mathematical methods. From the reference distribution, a range of acceptable statistic values is determined based on a preset coefficient $\gamma = 1 - \alpha$ (for example, $\gamma = 0.99$), where α is the significance level, i. e. γ is the probability that the statistic value is inside of the acceptable statistic value range. If the test statistic value lies outside the acceptable statistic value range, the hypothesis $\mathcal{H}_0$ is rejected, otherwise, the hypothesis $\mathcal{H}_0$ is accepted.

A randomness test suite may contain a serial of tests, which evaluate different properties of sequences. Since the ranges of acceptable values of various test statistics corresponding to the same γ are different, the p-values p, determined from the distributions of test statistics, may be used as unified numerical characteristics of the test results for sequences. In publication [2] the p-value is defined as «the probability that a perfect RNG would have produced a sequence less random than the sequence that was tested» [2, p. 1-4]. More specifically, the p-value is the probability of obtaining a statistic value s equal to or «more extreme» than the observed statistic value s_{obs} for the tested sequence, under the assumption that the null hypothesis is correct. According to the definition of «more extreme» cases, the tests are generally divided into two categories – one-sided tests and two-sided tests.

When $\mathcal{H}_0$ is true and $p \le \alpha$ we have a false positive in the test interpretation. This is called type I error, we can compute its probability since we have a complete characterisation of the sequences generated by a TRNG.

[3]Remind that the p-value is the probability to obtain the value of statistic at least as large as the actually observed one, under the assumption that the null hypothesis is true.

It is known that if the distribution of statistic is continuous then p-values should be uniformly distributed on the interval $[0,1]$ under the null hypothesis. Since p is uniformly distributed, the probability of a type I error is

$$\mathbf{P}_{\mathcal{H}_0}\{p \le \alpha\} = F_p^{(0)}(\alpha) = \alpha,$$

where $F_p^{(0)}$ is the distribution function of the random variable p having a uniform distribution on the interval $[0,1]$. For this reason, α is also called level of significance or statistical significance. The value of α is usually small, the significance level recommended by NIST is $\alpha = 0.01$.

On the contrary, the fact that $p > \alpha$ when $\mathcal{H}_0$ is false we have a false negative in the test interpretation. This is called type II error (accepting the sequence as random when $\mathcal{H}_0$ is false), and we denote this probability by β. The value $(1-\beta)$ is called the statistical power of the test, its exact computation is a difficult task, since it depends on the specific probabilistic mathematical model of the (non-ideal) generator under test.

So we can commit two errors:

- type I – reject $\mathcal{H}_0$, when the sequence is generated by an ideal RNG;
- type II – accept $\mathcal{H}_0$, when the sequence is generated by a RNG that is non-ideal.

It's important to find a balance between the probabilities of making type I and type II errors. Reducing α always comes at the cost of increasing β, and vice versa, with a fixed sample size. This approach is known as first-level testing and is followed by paper [4].

Second-level testing approach of the NIST SP 800-22

The usual way to test a TRNG or PRNG is to generate a sequence of n bits and analyse it with the test suite. Given a level of significance α, the sequence may be considered random if all tests in the suite produce p-values greater than α.

This approach presents a serious weakness. It is well known that some PRNGs can easily pass all tests. For example, a periodic (and thus, not ideal) generator always passes the Frequency test if the number of ones and zeros in the sequence is balanced.

To overcome this weakness, a more intensive test is necessary, involving a number N of different sequences generated by the RNG under test. In publication [2] NIST recommends using the second-level testing approach: a long binary sequence is partitioned into N subsequences, each with n bits. A standard test is applied for each sequence, and the distribution of the N obtained p-values is compared with a uniform distribution $F_p^{(0)}$. Note that from the definition of null hypothesis it follows that if $\mathcal{H}_0$ is true, then the N sequences (and so the N p-values) are independent of one another and approximately uniformly distributed on the interval $[0,1]$. To check this in publication [2] NIST proposes a chi-square goodness-of-fit test, this test is again a statistical test and gives another (a second-level) p-value p_{II}.

The NIST SP 800-22 suite includes 15 tests and for each of them the second-level test is performed. For each test and for each of N blocks of the sequence N p-values are computed, then the following second-level test is performed based on these N p-values.

This approach has been known for a long time [3] and it may increase the testing power as compared to a standard approach [4; 9].

1. Given the empirical results for a particular statistical test, compute the fraction ζ of the sequences passed the test. For example, if N binary sequences were tested with the significance level $\alpha = 0.01$ and m binary sequences had p-values equal to or greater than α, then the fraction is $\zeta = \frac{m}{N}$. If $\mathcal{H}_0$ is true, then the distribution of ζN is binomial with parameters N and $(1-\alpha)$; if N is large enough (e. g., $N \ge 1000$), then the distribution of ζ may be approximated by the distribution of the random variable having a normal distribution $\mathcal{N}(\mu, \sigma)$, where

$$\mu = 1 - \alpha;\ \sigma = \sqrt{\frac{\alpha(1-\alpha)}{N}}.$$

The range of acceptable fractions may be determined by the interval

$$\left[1 - \alpha - 3\sqrt{\frac{\alpha(1-\alpha)}{N}},\ 1 - \alpha + 3\sqrt{\frac{\alpha(1-\alpha)}{N}}\right].$$

If the ζ falls outside of this interval, then there is evidence that the data does not conform to the null hypothesis and $\mathcal{H}_0$ is rejected. The normal distribution approximation for the binomial distribution is reasonably accurate for large sample sizes.

2. Given N sequences, the distribution of p-values is examined to check the uniformity. The interval $[0,1]$ is divided into k disjoint subintervals, and the number of p-values falling in each subinterval are counted. Uniformity is determined via an application of a chi-square goodness-of-fit test in k subinterval with statistic χ^2. This is accomplished by computing chi-square statistic value

$$\chi^2_{\text{obs}} = \sum_{i=1}^{k} \frac{\left(\pi_i - \frac{N}{k}\right)^2}{\frac{N}{k}},$$

where π_i is the number of p-values in i^{th} subinterval. This statistical test yields a level-two p-value $p_{\text{II}} = \mathbf{P}_{\mathcal{H}_0}\left\{\chi^2 > \chi^2_{\text{obs}}\right\}$ (here χ^2 is a random variable having chi-square distribution with $(k-1)$ degrees of freedom), which is calculated as follows:

$$p_{\text{II}} = \frac{1}{\Gamma\left(\frac{k-1}{2}\right)} \int\limits_{\frac{\chi^2_{\text{obs}}}{2}}^{\infty} t^{\frac{k-1}{2}-1} e^{-t} dt,\ \Gamma\left(\frac{k-1}{2}\right) = \int\limits_{0}^{\infty} t^{\frac{k-1}{2}-1} e^{-t} dt.$$

Given a significance α_{II}, $\mathcal{H}_0$ is rejected if $p_{\text{II}} \le \alpha_{\text{II}}$, otherwise the sequences may be considered to be uniformly distributed. In publication [2] NIST recommends to use $\alpha_{\text{II}} = 0.0001$ and $k = 10$ subintervals.

The choice of N in a two-level approach is usually a trade-off. For example, the length of the sequence n may be limited by the memory size or by the computational power available. On the other hand, it may be preferable to test short sequences, thereby limiting n.

In every statistical test some approximations are adopted, introducing errors in the p-value computation and so in the p-value distribution. It was observed [4; 9] that for extremely large values of N, the level-two approach always fail, it ends with $p_{\text{II}} \simeq 0$. In this case we can say that the test is not reliable, since its significance is sensible different from the expected one. From this point of views, N is also a trade-off between statistical power $(1-\beta)$ and reliability of a level-two test.

Advantages of two-level test

The undoubted advantage of using the two-level test, for example, is rejection the generator with a small period size by Frequency test: regardless of the sequence length, a basic first-level test is always passed, while the second-level test is able to recognise a generator that always passes a basic test is not ideal.

Two PRNGs were considered in work [4]: the 32-bit version of the KISS [11], which is a very simple but effective generator, and the BBS generator [12]:

$$x_{n+1} = x_n^2 \bmod M,\ M = P \cdot Q,$$

where P and Q are large distinct primes; x_0 is a quadratic residue $\bmod M$; the output of the BBS generator is the bit parity of x_{n+1}, that is a computationally complex PRNG that, under the assumption of intractability for quadratic residuosity problem modulo M, has proven to be cryptographically secure (i. e. polynomial-time unpredictable (see [12])). For both generators in work [4] were performed a first-level test on a single sequence and a second-level test, checking the uniformity of $N = 10\,000$ p-values, on N different sequences with a chi-square test over 16 subintervals, and with a Kolmogorov – Smirnov test.

The empirical distributions of the tests of p-values were computed for both generators, which were then compared with the uniform distribution in the interval $[0,1]$ by calculating the p-values p_{II}. In this case, $\mathcal{H}_0$ correspond to «the two distributions match»; again, we reject $\mathcal{H}_0$ if $p_{\text{II}} \le \alpha_{\text{II}}$ and accept $\mathcal{H}_0$ if $p_{\text{II}} > \alpha_{\text{II}}$. Although in publication [2] NIST recommends to use $\alpha_{\text{II}} = 0.0001$, the value of α_{II} in work [4] was set to $\alpha = 0.01$ to make possible a direct comparison between a first-level test and a second-level test. The authors noted that the comparison may seem unfair, since a first level-test considers n bits, while a second-level test – nN bits.

Both generators had passed all first-level tests. However (except for the Discrete Fourier transform test that is well-known to have an error in the parameter of the reference distribution of the considered statistic), only the BBS generator passed the second-level tests. The proposed second-level test is able to recognise the non-randomness of the KISS generator, while a simple first-level test fails. In addition to the above experiment, the authors of work [4] considered a second one involving a much larger number of sequences generated with the BBS algorithm ($N = 150\,000$ sequences) in order to obtain more reliable results. The tests were also

applied to three high-end physical process based TRNGs. All three generators have been considered with an additional post processing stage, this in order to hide all possible imperfections and be sure that the properties of analysing streams are close to the properties of sequences of independent and equiprobable bits. Results of experiments are far from the desired ones, since too many tests fail, in particular, only 6 out of 15 two-level tests were passed by the BBS generator.

Experiments

All 15 tests from table 1 were applied to pseudorandom sequences generated by AES block cipher. The detailed description of the design of AES-based RNG may be found in papers [13; 14]. For generation such sequences AES block cipher in the CFB (*cipher feedback block*) mode (fig. 1) with the zero initialisation vector (IV) and plaintext block (Plaintext) was used. The key for each sequence was randomly selected from the set of 128-bit binary sequences, the length in bits of all sequences was chosen to be the same and equal to $n = 2^{20} = 1\,048\,576$ each (value corresponds to the NIST input size recommendation for the length of tested sequences).

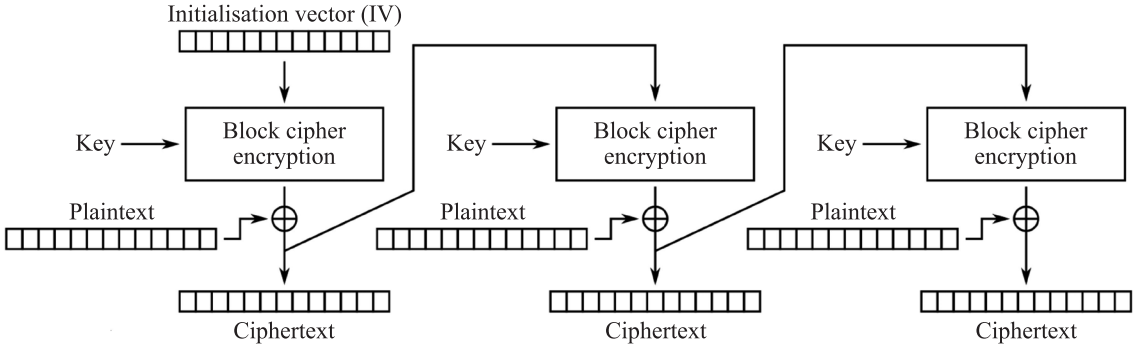

Fig. 1. CFB mode

Keys (on fig. 1) were obtained using the open source cryptographic library OpenSSL:

```
int RAND_bytes(unsigned char *buf, int num).
```

This function generates `num` bytes using the cryptographically secure PRNG (CSPRNG), and saves them in the `buf` array. By default, the OpenSSL CSPRNG supports a security level of 256 bits, provided it was able to seed itself from a trusted entropy source[4].

If we will use the AES-based PRNG, then this allow us to arbitrarily increase N. The effect of increasing the testing power can be seen in the example of table 2, where the testing results for the AES-based RNG are shown. The two-level test was performed for all 188 statistics values computed by NIST test suite, but only 15 of them are presented in table 2. In order to have the same significance in all tests, α_{II} in all experiments was set equal to $\alpha = 0.01$. These results confirm that for extreme values of N (e. g., $N = 2^{20}$) the two-level testing approach was carried out on the sequences obtained by AES algorithm (it is known that such sequences are practically indistinguishable from random one (see [13; 14])) are failed too, i. e. it ends with $p_{II} \simeq 0$.

Table 2

Results of the chi-square-based two-level randomness test for the AES-based RNG, with N ranging from 10^3 to 2^{20}

Number	Test name	Results of the chi-square test			
		$N = 10^3$	$N = 10^4$	$N = 10^5$	$N = 2^{20}$
1	Frequency test	0.616 305	0.290 806	0.588 411	**0.000 803**
2	Block frequency test	0.187 581	0.773 212	0.374 097	**0.000 125**
3	Cumulative sums test	0.401 199	0.124 765	0.959 543	**0.000 009**
4	Runs test	0.150 340	0.885 418	0.910 568	0.107 966
5	Longest run test	0.610 070	0.239 883	**0.000 355**	**0.000 000**

[4]RAND_bytes [Electronic resource]. URL: https://www.openssl.org/docs/man1.1.1/man3/RAND_bytes.html (date of access: 21.06.2025).

Ending of the table 2

Number	Test name	Results of the chi-square test			
		$N = 10^3$	$N = 10^4$	$N = 10^5$	$N = 2^{20}$
6	Binary matrix rank test	0.878 618	0.341 017	**0.000 000**	**0.000 000**
7	Discrete Fourier transform test	0.371 941	0.014 836	**0.000 000**	**0.000 000**
8	Overlapping template matching test	0.071 177	0.202 268	**0.000 000**	**0.000 000**
9	Maurer's universal statistical test	0.574 903	0.108 534	**0.000 000**	**0.000 000**
10	Approximate entropy test	0.246 750	0.078 038	0.219 501	**0.000 000**
11	Serial test	0.942 198	0.174 057	0.213 964	0.572 679
12	Linear complexity test	0.839 507	0.279 152	0.299 852	0.117 305
13	Non-overlapping template matching test	0.092 041	0.372 782	0.121 382	0.275 416
14	Random excursions test	0.914 727	0.663 838	0.346 173	**0.000 028**
15	Random excursions variant test	0.238 264	0.133 576	**0.000 080**	**0.000 000**

Note. Tests, where $p_{\text{II}} \le 0.01$, are in bold.

In order to identify the problem, let's take a look on the simple Frequency test.

Frequency test

The purpose of the Frequency test is to determine whether the number of ones and zeros in a sequence are approximately the same as would be expected for a truly random sequence. The test assesses the closeness of the fraction of ones to $\frac{1}{2}$, that is, the number of ones and zeros in a sequence should be about the same. All subsequent tests depend on passing this first basic test (see [2, p. 2-2]).

Let n be the length of the bit string, $\varepsilon = \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ be the input sequence of bits and the null hypothesis $\mathcal{H}_0$ is that the sequence ε consists of independent identically distributed Bernoulli random variables (X_i or ε_i), where $X_i = 2\varepsilon_i - 1$, and so the probability of ones is $\frac{1}{2}$. Denote by

$$S_n = X_1 + \ldots + X_n = 2(\varepsilon_1 + \ldots + \varepsilon_n) - n.$$

Under the null hypothesis, S_n is assumed to follow a binomial distribution. By the classic de Moivre – Laplace theorem, for a sufficiently large number of trials, the distribution of the binomial sum, normalised by $\sqrt{n}$, is closely approximated by a standard normal distribution. This test makes use of that approximation to assess the closeness of the fraction of ones to $\frac{1}{2}$.

The test is derived from the well-known central limit theorem for the random walk, $S_n = X_1 + \ldots + X_n$. According to the central limit theorem,

$$\lim_{n\to\infty} F_{S_n}(z) = \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{\frac{-u^2}{2}} du,$$

where $F_{S_n}(z) = \mathbf{P}\left\{\frac{S_n}{\sqrt{n}} < z\right\}$. This classical result serves as the basis of the simplest test for randomness. It implies that, if S_n is normal, then $|S_n|$ is half normal (i. e. $F_{|S_n|}(z) = 2F_{S_n}(z)$, $z \ge 0$) and

$$F_{|S_n|}(z) = \mathbf{P}\left\{\frac{|S_n|}{\sqrt{n}} \le z\right\}, \quad \lim_{n\to\infty} F_{|S_n|}(z) = 2\Phi(z) - 1.$$

So, according to the test statistic $S = \frac{|S_n|}{\sqrt{n}}$, it is necessary to determine observed value $S(\varepsilon) = \frac{|X_1 + \ldots + X_n|}{\sqrt{n}}$, and then calculate the corresponding p-value, which is equal to

$$p_{\mathrm{I}} = \lim_{n\to\infty}\left(1 - F_{|S_n|}\left(S(\varepsilon)\right)\right) = 2\left(1 - \Phi\left(S(\varepsilon)\right)\right) = \operatorname{erfc}\left(\frac{S(\varepsilon)}{\sqrt{2}}\right),$$

where $\operatorname{erfc}(\cdot)$ is the complementary error function

$$\operatorname{erfc}(z) = \frac{2}{\sqrt{\pi}}\int_z^\infty e^{-u^2}\,du.$$

In fig. 2, *a*, is shown the histogram of the set of *p*-values for $k = 10$ subintervals corresponding to the Frequency test applied to the AES-based generator, and for the comparison in fig. 2, *b*, – the Binary matrix rank test from the NIST test suite applied to the same sequences from the AES-based generator. If we consider the theoretical standard deviation of the number of random $N = 2^{20}$ independent, uniformly distributed values over $k = 10$ subintervals, we easily get $\sigma = \dfrac{\sqrt{N(k-1)}}{k}$. In this case, $\sigma \simeq 307.2$, the upper and lower continuous lines in the figures correspond to the deviation of 3σ from the mean value; the figure shows that the observed deviation is far from this value.

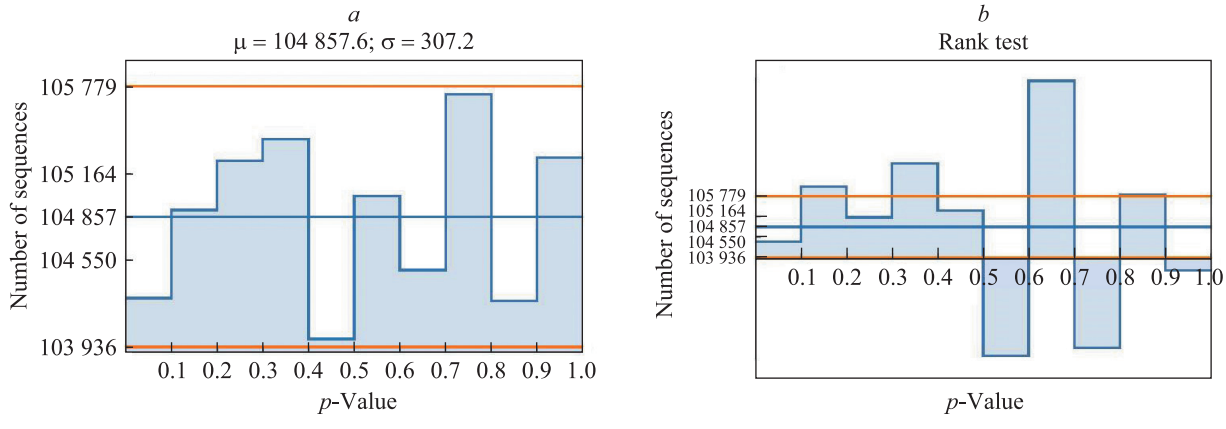

Fig. 2. Comparison between expected deviation and measured deviation in the distribution of *p*-values in the interval $[0,1]$ for the Frequency test (*a*) and the Binary matrix rank test (*b*)

Let $p_1, \ldots, p_N \in [0,1]$ be the sample of *p*-values with assumed continuous CDF $F_p^{(0)}(x)$. The empirical CDF $F_N(x)$ for this sample is defined as follows:

$$F_N(x) = \frac{1}{N}\sum_{j=1}^{N} H(x - p_j),$$

where $H(x)$ is the unity-step function, defined as $H(x) = \begin{cases} 1, & x \ge 0, \\ 0, & x < 0. \end{cases}$ The Kolmogorov – Smirnov statistic for a given CDF $F_p^{(0)}(x)$ is

$$D_N = \sup_x \left| F_N(x) - F_p^{(0)}(x) \right|.$$

Let $\mathcal{K}$ be the CDF of D_N, then the *p*-values is given by $1 - \mathcal{K}(D_N)$. Under null hypothesis that the sample comes from the hypothesised distribution $F(x)$

$$\sqrt{N} D_N \xrightarrow[N\to\infty]{} \sup_t \left| B\left(F_p^{(0)}(t)\right)\right|$$

in distribution, where $B(t)$ is the Brownian bridge.

In fig. 3, *a*, is shown the discrete empirical CDF F_N of *p*-values for the Frequency test applied to the AES-based sequences, in fig. 3, *b*, are presented the differences between empirical CDF F_N and uniform CDF, $n = N = 2^{20}$, from where it is clear that the modulus of the difference between F_N and $F_p^{(0)}$ does not exceed 0.001 2.

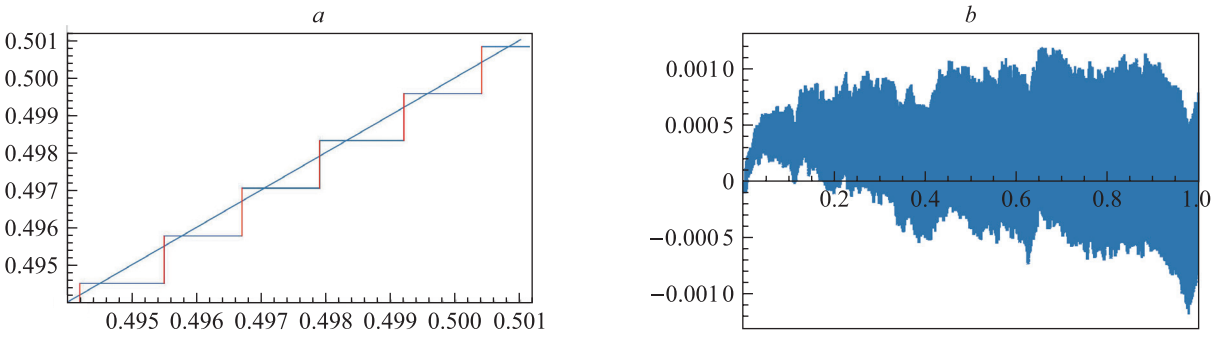

Fig. 3. Comparison between empirical CDF F_N of the p-values generated by Frequency test for $n = 2^{20}$ and uniform CDF (*a*). Deviation of the empirical CDF F_N of the p-value distribution of the Frequency test from continuous uniform CDF (*b*)

The deviation may be identified with an error propagated from the computation of the p-value in the first-level test, which is a consequence of the introduced approximations. The random variable S_n in relation (1) has a binomial distribution, which is however assumed to be normal. From Berry – Esseen theorem, we know that the error of this approximation, under the assumptions of the central limit theorem, is bounded by (see [15])

$$\sup_x \left| F_{S_n}(x) - \Phi(x) \right| \le \frac{C\mathbf{E}\left|X_i\right|^3}{\sigma^3\sqrt{n}},$$

where $F_{S_n}(z) = \mathbf{P}\left\{\frac{S_n}{\sqrt{n}} < z\right\}$; $\Phi(\cdot)$ is a CDF of the standard normal distribution; n is the number of independent variables $X_i, 1 \le i \le n$, summed in relation (1), i. e. the number of bits in the sequence; $\sigma = \mathbf{E}\left(X_i^2\right) = 1$; the third order moment is $\mathbf{E}\left|X_i\right|^3 = 1$; $C < 0.474\,8$ (see [16]). The maximal error ϵ of p-value is the error of the approximation $F_{|S_n|}(x)$ by $\Phi(x)$ and from formula (2) is twice the above error, i. e. $\epsilon = \frac{2C}{\sqrt{n}}$. If $n = 2^{20}$, then $\epsilon = 9.3 \cdot 10^{-4}$.

Assuming this bound of the error in the computation of a p-value, we can bound also the maximal error in the distribution of N p-values in k subintervals. A p-value that should belong to a subinterval may be found into the nearby one only if its distance from the border of two subintervals is less than ϵ. If we have N p-values uniformly distributed in the interval $[0,1]$ the maximal number of p-values that may be found in the wrong adjacent subinterval is ϵN. This is independent of the numbers of subintervals.

Since all subintervals (but the first and the last), have two neighbours, the maximum error Δ in the number of p-values in a subinterval is $\Delta = 2N\epsilon$. In our case, $N = 2^{20}$, so $\Delta \simeq 1944.8$. This value is compatible with that may be observe in fig. 2, *a*. The increasing n will result in decreasing the propagated error as $\frac{1}{\sqrt{n}}$ (the results of the experiments are confirm that in paper [4]). More generally, increasing n will reduce the error in the first-level p-value and so increasing the reliability of a second-level test, for all tests in the suite. A very simple reliability condition is requiring that Δ is smaller than the standard deviation of the number of p-values in a subinterval from its mean value, i. e.

$$\Delta < \frac{\sqrt{N(k-1)}}{k}.$$

In the case $\Delta = 2N\epsilon$, $k = 10$ and $n = 2^{20}$, we get

$$N \le \frac{1}{4\epsilon^2 k}\left(1 - \frac{1}{k}\right) \simeq 26\,014.6.$$

Two-sided estimates for the quantiles of the binomial distribution

To construct a statistical test with the specified error probabilities, we need the information on the distribution «tails». But relative errors of the de Moivre – Laplace approximations for the tails of binomial distribution

function are large. This was one of the first problems in the theory of large deviations. There are a lot of inequalities for the tails of binomial distribution function, as well as for the distributions of sums of random variables.

In paper [17] explicit two-sided estimates for the distribution function $F_{n,p}(k)$ of the binomial law with parameters n, p were obtained. These results are close in form to the Bernstein and Feller inequalities, but have an explicit form and complement them, almost completely solving the problem of estimating the probabilities of moderate and large deviations for binomial laws for any values of the parameters.

Let $X_{n,p}$ be a random variable with the binomial distribution with parameters n, p:

$$\mathbf{P}\left\{X_{n,p} \le k\right\} = \sum_{0 \le m \le k} C_n^m p^m (1-p)^{n-m}.$$

Theorem 1 [17]. *Let*

$$H(x, p) = x \ln\left(\frac{x}{p}\right) + (1-x)\ln\left(\frac{1-x}{1-p}\right), \ \operatorname{sign}(x) = \frac{x}{|x|}$$

for $x \ne 0$ *and* $\operatorname{sign}(0) = 0$, *and let* $\left\{C_{n,p}(k)\right\}_{k=0}^{n}$ *be increasing sequences defined as follows:*

$$C_{n,p}(0) = (1-p)^n, \ C_{n,p}(n) = 1 - p^n,$$

$$C_{n,p}(k) = \Phi\left(\operatorname{sign}\left(\frac{k}{n} - p\right)\sqrt{2nH\left(\frac{k}{n}, p\right)}\right), \ 1 \le k < n.$$

Then for every $k = 0, 1, \ldots, n-1$ *and for every* $p \in (0,1)$

$$C_{n,p}(k) \le \mathbf{P}\left\{X_{n,p} \le k\right\} \le C_{n,p}(k+1)$$

and equalities hold only for $k = 0$ *and* $k = n-1$.

We will use the following standard notations:

$$\lfloor x \rfloor = \max\left\{m \in \mathbb{Z} \mid m \le x\right\}, \ \lceil x \rceil = \min\left\{n \in \mathbb{Z} \mid n \ge x\right\}.$$

By definition of the α-level quantile $x_\alpha(n, p)$ of the binomial distribution with parameters n, p

$$x_\alpha(n, p) = \min\left\{k : \mathbf{P}\left\{X_{n,p} \le k\right\} \ge \alpha\right\} \in \{0, 1, \ldots, n\},$$

$$x_0(n, p) = 0 \ \text{ and } \ x_1(n, p) = n,$$

$$x_{\frac{1}{2}}\left(n, \frac{1}{2}\right) = \left\lfloor \frac{n}{2} \right\rfloor.$$

From theorem 1 for any $\alpha \in [0,1]$ it follows that

$$\min\left\{k : C_{n,p}(k+1) \ge \alpha\right\} \le x_\alpha(n, p) = \min\left\{k : \mathbf{P}\left\{X_{n,p} \le k\right\} \ge \alpha\right\} \le \min\left\{k : C_{n,p}(k) \ge \alpha\right\}.$$

Recently the author had obtained the two-sided estimates for the α-level quantile $x_\alpha(n, p)$ of the binomial law. Such estimates allow for any α-level and any values of binomial law parameters n, p to determine a narrow integer interval containing the quantile $x_\alpha(n, p)$, it has been experimentally shown that quite often the length of this interval is one, i. e. the interval boundaries are two subsequent integers one of which is the quantile $x_\alpha(n, p)$. These estimates allow to completely eliminate the systematic error in determining one of the $k = 10$ disjoint subintervals on the interval $[0,1]$ for the Frequency test to which the p-value belongs.

Theorem 2 [18]. *The following estimates are true:*

- *for* $0 < \alpha < \frac{1}{2}$

$$r_1 - 1 \le x_\alpha(n, p) \le r_2,$$

where $q = 1 - p$, $r_1 = \left\lceil np + \frac{q-p}{6}\Phi^{-1}(\alpha)^2 + \Phi^{-1}(\alpha)\sqrt{npq + \frac{(q-p)^2\Phi^{-1}(\alpha)^2}{6^2}}\right\rceil$, and

$$r_2 = \left\lceil \frac{(6-5p)np + \frac{q}{2}\Phi^{-1}(1-\alpha)\left(3q\Phi^{-1}(1-\alpha) - \sqrt{24(6-5p)np + (3+p)^2\Phi^{-1}(1-\alpha)^2}\right)}{(6-5p) + \frac{2q\Phi^{-1}(1-\alpha)^2}{n}}\right\rceil;$$

- *for* $\frac{1}{2} < \alpha \le 1$

$$r_3 \le x_\alpha(n, p) \le r_1,$$

where $r_3 = \left\lceil \frac{(1+5p)np + \frac{p}{2}\Phi^{-1}(\alpha)\left((4-3p)\Phi^{-1}(\alpha) + \sqrt{24n(1+5p)q + \Phi^{-1}(\alpha)^2(4-p)^2}\right)}{(1+5p) + \frac{2p\Phi^{-1}(\alpha)^2}{n}}\right\rceil - 1;$

- *for* $\alpha = \frac{1}{2}$

$$x_{\frac{1}{2}}(n, p) = \begin{cases} \lfloor np \rfloor \text{ or } \lceil np \rceil, \text{ if } p \ne \frac{1}{2}, \\ \frac{n-1}{2}, \text{ if } p = \frac{1}{2}, n - odd, \\ \frac{n}{2}, \text{ if } p = \frac{1}{2}, n - even. \end{cases}$$

Corollary. *If* $0 < \alpha < \frac{1}{2}$, *then the following estimates are true:*

$$\left\lceil \frac{n}{2} + \frac{\sqrt{n}}{2}\Phi^{-1}(\alpha)\right\rceil - 1 \le x_\alpha\left(n, \frac{1}{2}\right) \le \left\lceil \frac{\frac{7n}{4} + \frac{3}{8}\Phi^{-1}(\alpha)^2 + \frac{1}{4}\Phi^{-1}(\alpha)\sqrt{42n + \left(\frac{7}{2}\right)^2\Phi^{-1}(\alpha)^2}}{\frac{7}{2} + \frac{\Phi^{-1}(\alpha)^2}{n}}\right\rceil,$$

if $\frac{1}{2} < \alpha < 1$, *then*

$$\left\lceil \frac{\frac{7n}{4} + \frac{5}{8}\Phi^{-1}(\alpha)^2 + \frac{1}{4}\Phi^{-1}(\alpha)\sqrt{42n + \left(\frac{7}{2}\right)^2\Phi^{-1}(\alpha)^2}}{\frac{7}{2} + \frac{\Phi^{-1}(\alpha)^2}{n}}\right\rceil - 1 \le x_\alpha\left(n, \frac{1}{2}\right) \le \left\lceil \frac{n}{2} + \frac{\sqrt{n}}{2}\Phi^{-1}(\alpha)\right\rceil,$$

if $\alpha = \frac{1}{2}$, *then*

$$x_{\frac{1}{2}}\left(n, \frac{1}{2}\right) = \begin{cases} \frac{n-1}{2}, \text{ if } n - odd, \\ \frac{n}{2}, \text{ if } n - even. \end{cases}$$

The 11 quantiles of binomial distribution with parameters $\left(2^{20}, \frac{1}{2}\right)$ are shown in table 3.

Table 3

Quantiles of binomial distribution

α	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$x_\alpha\left(2^{20}, \frac{1}{2}\right)$	0	523 632	523 857	524 020	524 158	524 288	524 418	524 556	524 719	524 944	2^{20}

Conclusions

The two-level approach for testing RNGs involving the well-known NIST SP 800-22 test suite was considered. Such approach may increase the reliability of the test. However, it is sensitive to the approximation error introduced by the computing of *p*-values. In this paper it has been shown that for AES-based sequences two-level testing approach is not reliable too. Systematic error in the computing of the *p*-values is dependent only on *n*, that is the number of bits in the analysed sequences. For a reliable second-level test, this error, based on the carried out heuristic considerations, should be no greater than $\frac{\sigma}{N} = \frac{1}{k}\sqrt{\frac{k-1}{N}}$, where σ is the standard deviation from the mean number of particles in a bin for equiprobable scheme of allocation *N* particles in *k* bins (disjoint subintervals of the interval $[0,1]$). For example, in the second-level test of the Frequency test with $n = 2^{20}$ the number of tested sequences *N* should not exceed 26 000.

To completely eliminate the systematic error appearing in the Frequency test when determining the number of bin from *k* bins to which the *p*-value belongs the two-sided estimates of the quantiles of the binomial law are proposed.

References

1. Menezes AJ, van Oorschot PC, Vanstone SA. *Handbook of applied cryptography*. Boca Raton: CRC Press; 1996. XXVIII, 780 p. (Discrete mathematics and its applications).

2. Rukhin A, Soto J, Nechvatal J, Smid M, Barker E, Leigh S, et al. *A statistical test suite for random and pseudorandom number generators for cryptographic applications.* Gaithersburg: National Institute of Standards and Technology; 2010. [131] p. (NIST special publications; SP 800-22 revision 1a).

3. L'Ecuyer P, Simard R. *TestU01: a software library in ANSI C for empirical testing of random number generators* [Internet]. Montreal: Université de Montréal; 2013 [cited 2025 June 21]. IV, 214 p. Available from: https://scispace.com/pdf/a-software-library-in-ansi-c-for-empirical-testing-of-random-hk10yxs2w3.pdf.

4. Pareschi F, Rovatti R, Setti G. Second-level NIST randomness tests for improving test reliability. In: Institute of Electrical and Electronics Engineers. *2007 IEEE International symposium on circuits and systems (ISCAS); 2007 May 27–30; New Orleans, USA*. [S. l.]: Institute of Electrical and Electronics Engineers; 2007. p. 1437–1440. DOI: 10.1109/ISCAS.2007.378572.

5. Kim S-J, Umeno K, Hasegawa A. *Corrections of the NIST statistical test suite for randomness*. arXiv:nlin/0401040v1 [Preprint]. 2004 [cited 2025 June 21]: [14 p.]. Available from: https://arxiv.org/abs/nlin/0401040v1.

6. Hamano K. The distribution of the spectrum for the Discrete Fourier transform test included in SP 800-22. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*. 2005;E88-A(1):67–73.

7. Hamano K, Kaneko T. Correction of overlapping template matching test included in NIST randomness test suite. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*. 2007;E90-A(9):1788–1792.

8. Sulak F, Doğanaksoy A, Ege B, Koçak O. Evaluation of randomness test results for short sequences. In: Carlet C, Pott A, editors. *Sequences and their applications – SETA 2010. Proceedings of the 6th International conference; 2010 September 13–17; Paris, France.* Berlin: Springer; 2010. p. 309–319 (Lecture notes in computer science; volume 6338). DOI: 10.1007/978-3-642-15874-2_27.

9. Pareschi F, Rovatti R, Setti G. On statistical tests for randomness included in the NIST SP 800-22 test suite and based on the binomial distribution. *IEEE Transactions on Information Forensics and Security*. 2012;7(2):491–505. DOI: 10.1109/TIFS.2012.2185227.

10. Kharin YuS, Zubkov AM. On statistical testing of composite hypotheses on *s*-dimensional uniform probability distribution of binary sequences. *Diskretnaya matematika*. 2024;36(1):116–135. Russian. DOI: 10.4213/dm1806.

11. Rose GG. KISS: a bit too simple. *Cryptography and Communications.* 2018;10(1):123–137. DOI: 10.1007/s12095-017-0225-x.

12. Blum L, Blum M, Shub M. A simple unpredictable pseudo-random number generator. *SIAM Journal on Computing*. 1986; 15(2):364–383. DOI: 10.1137/0215025.

13. Zubkov AM, Serov AA. Testing the NIST statistical test suite on artificial pseudorandom sequences. *Matematicheskie voprosy kriptografii*. 2019;10(2):89–96. DOI: 10.4213/mvk286.

14. Zubkov AM, Serov AA. A natural approach to the experimental study of dependence between statistical tests. *Matematicheskie voprosy kriptografii*. 2021;12(1):131–142. DOI: 10.4213/mvk352.

15. Shiryaev AN. *Probability*. 2nd edition. Boas RP, translator. Berlin: Springer-Verlag; 1996. XVI, 623 p. (Graduate texts in mathematics; volume 95).

16. Korolev V, Shevtsova I. An improvement of the Berry – Esseen inequality with applications to Poisson and mixed Poisson random sums. *Scandinavian Actuarial Journal*. 2012;2012(2):81–105. DOI: 10.1080/03461238.2010.485370.

17. Zubkov AM, Serov AA. A complete proof of universal inequalities for distribution function of binomial law. *Theory of Probability and its Applications*. 2013;57(3):539–544. DOI: 10.1137/S0040585X97986138.

18. Serov AA. Two-sided estimates of the binomial law quantiles. *Theory of Probability and its Applications*. Forthcoming 2026.

Received 24.12.2025 / revised 19.01.2026 / accepted 29.01.2026.