Autocorrelation Criterion for Quality Assessment of Random Number Sequences

Emil Faure[1]✉ [0000-0002-2046-481X] , Iryna Myronets[2] [0000-0003-2007-9943], Artem Lavdanskyi[3] [0000-0002-1596-4123]

Cherkasy State Technological University, Shevchenko Blvd., 460, Cherkasy, 18006, Ukraine

[1]e.faure@chdtu.edu.ua
[2]i.myronets@chdtu.edu.ua
[3]a.lavdanskyi@chdtu.edu.ua

Abstract. The authors analyze different approaches to forming estimates of autocorrelation coefficients of random and pseudorandom number sequences. An integral estimate of normalized autocorrelation coefficients is theoretically obtained. Estimates of some statistical properties of normalized autocorrelation coefficients have been improved. The autocorrelation criterion for quality assessment of time series based on simultaneous analysis of several autocorrelation coefficients has been further developed by adapting it to uniformly distributed random variables. The technique of its implementation is presented. Applying the criterion revealed statistical deviations for some pseudorandom number generators that successfully pass all TestU01 autocorrelation tests.

Keywords: Random numbers · Random number generator · Time series · Correlation · Autocorrelation · Quality assessment · Statistical criterion.

1 Introduction

Since random number sequences are multi-parameter processes, different methods and criteria are used for assessing their quality. These methods and criteria consider random processes from different standpoints using different statistical estimates.

The most famous test suits are: the Donald Knuth's statistical tests set [1]; George Marsaglia's DIEHARD tests [2]; NIST Statistical Test Suite [3]; TestU01[4]. In addition, there are other packages and tests. Among them CRYPT-X [5], NIST PUB FIPS 140-2 [6] can be distinguished.

Based on the definition of discrete white noise [7], an autocorrelation test is one of the most common tests, which allow detecting statistical irregularities of the studied sequences of numbers.

In [1], as in [8], it is recommended to use the criterion of serial (cyclic) correlation between cyclically shifted copies of the studied sequence. In [4, 9, 10], estimates of autocorrelation coefficients are formed by comparing two subsequences. The count of the subsequences' elements is made from the beginning and from the end of the studied sequence. In [11], the autocorrelation coefficients are calculated for sequence successive intervals, which can overlap or stand away from each other.

Thus, there are different approaches to finding empirical autocorrelation coefficients that form an autocorrelation function (ACF) estimate. However, the main task of the autocorrelation test is to determine the correspondence of the sequence ACF estimate to the ACF of random number sequence described by the Dirac delta function [12]. Thus, estimates of the autocorrelation coefficients at non-zero points should go to zero. Their significance is most often evaluated by the Student's criterion [13].

The authors of the work [4] verify the correspondence of the distribution of autocorrelation coefficients estimates at nonzero points to the binomial law. For a large number of values, it can be approximated to normal.

In [14, 15] the statistical criteria of ACF side lobes complex estimate are proposed. Instead of testing the significance of each individual autocorrelation coefficient, these criteria check several autocorrelation coefficients to be different from zero. However, the proposed criteria are not adapted to analyze sequences of uniformly distributed random and pseudorandom numbers.

Thus, the question of estimating autocorrelation coefficients needs further study. This leads to the need for deeper analysis to identify the correlation properties inherent in sequences generated by natural sources of discrete white noise and not inherent in artificially generated pseudorandom sequences (PRS).

The purpose of this work is to develop a criterion for assessing the quality of sequences of uniformly distributed random and pseudorandom numbers, which allows detecting statistical irregularities not detected to date.

2 Formal Problem Statement

According to [16], the normalized autocorrelation coefficient $\rho_x\left(t',t''\right)$ of a random process $X\left(t\right)$ is calculated in the general form according to the expression:

$$\rho_X\left(t',t''\right)=K_X\left(t',t''\right)\Big/\left(\sqrt{Var\left[X\left(t'\right)\right]}\cdot\sqrt{Var\left[X\left(t''\right)\right]}\right),$$

where $K_X\left(t',t''\right)=E\left[\left(X\left(t'\right)-E\left[X\left(t'\right)\right]\right)\cdot\left(X\left(t''\right)-E\left[X\left(t''\right)\right]\right)\right]$ is the correlation moment (covariance coefficient) of the intersections $X\left(t'\right)$ and $X\left(t''\right)$ of the random process $X\left(t\right)$ at times t' and $t''=t'+\tau$; $E\left[X\left(t'\right)\right]$, $E\left[X\left(t''\right)\right]$, $Var\left[X\left(t'\right)\right]$, $Var\left[X\left(t''\right)\right]$ are the expectations and the variances of $X\left(t'\right)$, $X\left(t''\right)$. We will represent the intersections $X\left(t'\right)$ and $X\left(t''\right)$ as random variables (r.v.) X' and X''. Then $E\left[X\left(t'\right)\right]=E\left(X'\right)$, $E\left[X\left(t''\right)\right]=E\left(X''\right)$, $Var\left[X\left(t'\right)\right]=Var\left(X'\right)$, $Var\left[X\left(t''\right)\right]=Var\left(X''\right)$.

Suppose that random number generator (RNG) or pseudorandom number generator (PRNG) forms a stationary random process. For such a process, $E\left(X'\right)=E\left(X''\right)=m_X=const$, $Var\left(X'\right)=Var\left(X''\right)=\sigma_X^2=const$, and an autocorrelation coefficient depends only on the lag $\tau=t''-t'$: $K_X\left(t',t''\right)=k_X\left(\tau\right)$.

Let $\overset{\square}{X}'=X'-E\left(X'\right)$, $\overset{\square}{X}''=X''-E\left(X''\right)$ be r.v. with expectations $E\left(\overset{\square}{X}'\right)=E\left(\overset{\square}{X}''\right)=0$ and variances $Var\left(\overset{\square}{X}'\right)=Var\left(\overset{\square}{X}''\right)=\sigma_X^2$. Then the correlation moment $k_X\left(\tau\right)=E\left(\overset{\square}{X}'\cdot\overset{\square}{X}''\right)$ and the normalized autocorrelation coefficient $\rho_X\left(\tau\right)=k_X\left(\tau\right)\Big/\left(\sqrt{Var\left(X'\right)}\cdot\sqrt{Var\left(X''\right)}\right)=E\left(\overset{\square}{X}'\cdot\overset{\square}{X}''\right)\Big/\sigma_X^2$.

Suppose that the r.v. $\overset{\square}{X}'$ and $\overset{\square}{X}''$ are uncorrelated (this corresponds to the properties of white noise intersections) and independent. Then the $\rho_X\left(\tau\right)$ value is invariant to the value of $\tau\neq0$ and $\rho_X\left(\tau\neq0\right)\equiv0$ [9].

The product of r.v. $\xi=\overset{\square}{X}'\cdot\overset{\square}{X}''$ can be considered as a r.v. with parameters: $E\left(\xi\right)=E\left(\overset{\square}{X}'\right)\cdot E\left(\overset{\square}{X}''\right)=0$ and $Var\left(\xi\right)=E\left(\xi^2\right)-E^2\left(\xi\right)=E\left[\left(\overset{\square}{X}'\cdot\overset{\square}{X}''\right)^2\right]=$ $=E\left(\overset{\square}{X}'^{\,2}\right)\cdot E\left(\overset{\square}{X}''^{\,2}\right)=Var\left(\overset{\square}{X}'\right)\cdot Var\left(\overset{\square}{X}''\right)$ or $Var\left(\xi\right)=\sigma_X^4$, and $\sigma_\xi=\sigma_X^2$. In this case $\rho_X\left(\tau\right)=E\left(\xi\right)/\sigma_\xi$.

If values of a discrete random process $\xi\left(t\right)$ form a countable set of independent values $\xi_1,\ \xi_2,\ldots,\xi_n,\ldots$ then $E\left(\xi\right)=\lim\limits_{n\to\infty}\left(1/n\sum_{i=1}^n\xi_i\right)$.

According to the Lindeberg-Levy theorem [16], if mutually independent r.v. $\xi_1,\ \xi_2,\ldots,\xi_n,\ldots$ are equally distributed and have an expectation $E\left(\xi\right)=a$ and a variance σ_ξ^2, then the value $\left(\sum_{i=1}^n\xi_i-na\right)\Big/\sigma_\xi\sqrt{n}$ is normally distributed when $n\to\infty$: $\left(\sum_{i=1}^n\xi_i-na\right)\Big/\sigma_\xi\sqrt{n}\to N\left(0;1\right)$. Given that $E\left(\xi\right)=a=0$,

$$\left(\sum_{i=1}^n\xi_i-na\right)\Big/\sigma_\xi\sqrt{n}=\sqrt{n}\cdot1/n\cdot\sum_{i=1}^n\xi_i\Big/\sigma_\xi=\sqrt{n}\,E\left(\xi\right)/\sigma_\xi=\sqrt{n}\rho_X\left(\tau\right)\to N\left(0;1\right).$$

In other words, the marginal distribution of the normalized autocorrelation coefficient $\rho_X\left(\tau\right)$ of a random process whose intersections are independent r.v., is a normal distribution with expectation $E\left(\rho_X\left(\tau\right)\right)=0$ and variance $Var\left(\rho_X\left(\tau\right)\right)=1/n$.

Consider that $Var\left(\rho_X(\tau)\right)=E\left(\rho_X^2(\tau)\right)-\left(E\left(\rho_X(\tau)\right)\right)^2=\lim\limits_{\mathrm{T}\to\infty}\left(1/\mathrm{T}\cdot\sum_{i=1}^{\mathrm{T}}\rho_X^2(\tau)\right)$. It follows that

$$\lim_{\substack{n\to\infty\\ \mathrm{T}\to\infty}}\left(\sum_{\tau=1}^{\mathrm{T}}\rho_X^2(\tau)\right)=\mathrm{T}/n\,. \qquad (1)$$

Expression (1) defines the limit value of the ACF side lobes power $W(\mathrm{T})=\sum_{\tau=1}^{\mathrm{T}}\rho_X^2(\tau)$: $\lim\limits_{\substack{n\to\infty\\ \mathrm{T}\to\infty}}\left(W(\mathrm{T})\right)=\mathrm{T}/n$.

In addition, since the r.v. $\rho_X(\tau)\to N(0;1/n)$ when $n\to\infty$, the r.v. $\sum_{\tau=1}^{\mathrm{T}}n\rho_X^2(\tau)$ has a distribution χ_{T}^2 with T degrees of freedom $(n\to\infty)$:

$$n\sum_{\tau=1}^{\mathrm{T}}\rho_X^2(\tau)\to\chi_{\mathrm{T}}^2\,. \qquad (2)$$

Then for side lobes power $nW(\mathrm{T})\to\chi_{\mathrm{T}}^2$.

Thus, expression (2) is an integral estimate for autocorrelation coefficients. It creates the preconditions for constructing statistical criteria for checking the correlation properties of sequences of random and pseudorandom numbers by their empirical estimates.

Normalized autocorrelation coefficients are strictly defined by the theoretical expression (1). However, the correlation properties estimate of empirical number sequence can significantly depend on the studied sequence properties and the experiment conditions. In particular, the sequence correlation properties estimate can be performed:

- on a sequence period in the case of its periodicity, by analyzing a periodic ACF (PACF);
- on some fixed size sample;
- in real time, when sequence elements arrive to the analyzer sequentially.

In addition, the estimating correlation properties of random and pseudorandom number sequences may be conducted under conditions where the distribution law of a discrete random variable (d.r.v.) and its parameters are either fully known or empirically determined. In all these cases, it is necessary to know the first and second initial moments of normalized autocorrelation coefficients for their integral estimate.

We estimate these statistical properties of the normalized autocorrelation coefficients of a discrete random process calculated according to the defined approaches.

3 Estimate of Statistical Properties of Normalized Autocorrelation Coefficients

3.1 Periodic ACF Estimate

ACF is periodic if an original sequence is also periodic. Moreover, as it shown in [17], it is advisable for periodic signals to estimate the probabilistic moments in the minimum period. Taking into account the symmetry property of the PACF graph with respect to the axis, which is half period from the y-axis, it is allowed to reduce the number of calculated values by 2 times.

Let the number sequence $\left(x_0, x_1, \cdots, x_{n-1}\right)$ be repeated periodically with period n. In this case, as shown in [1] and [8], the estimate of the normalized autocorrelation coefficient is calculated according to the expression:

$$r_{PACF\ x}(\tau)=\frac{\sum_{i=0}^{n-1}\left[\left(x_i-\overline{x}\right)\cdot\left(x_{(i+\tau)\bmod n}-\overline{x}\right)\right]}{\sum_{i=0}^{n-1}\left(x_i-\overline{x}\right)^2}=\frac{n\sum_{i=0}^{n-1}x_i x_{(i+\tau)\bmod n}-\left(\sum_{i=0}^{n-1}x_i\right)^2}{n\sum_{i=0}^{n-1}x_i^2-\left(\sum_{i=0}^{n-1}x_i\right)^2}, \qquad (3)$$

where $\overline{x}=\sum_{i=0}^{n-1}x_i\Big/n$ is a statistical estimate of the X expectation.

Note that the estimate (3) is performed for the whole sequence period. Therefore, the estimate of the d.r.v. X expectation coincides with the expectation: $\overline{x}=\sum_{i=0}^{n-1}x_i\Big/n=\lim_{m\to\infty}\left(\overline{x}=\sum_{i=0}^{m}x_i\Big/m\right)=E(X)$. Based on similar considerations, the variance estimate over the entire sequence period also coincides with the d.r.v. variance: $\sigma_x^2=\sum_{i=0}^{n-1}\left(x_i-\overline{x}\right)^2\Big/n=\lim_{m\to\infty}\left(\sum_{i=0}^{m}\left(x_i-E(X)\right)^2\Big/m\right)=Var(X)=\sigma_X^2$. Then the expression (3) can be represented as $r_{PACF\ x}(\tau)=1/n\cdot\sum_{i=0}^{n-1}\left[\left(x_i-E(X)\right)\cdot\left(x_{(i+\tau)\bmod n}-E(X)\right)\right]\Big/\sigma_X^2$.

Obviously, given the symmetry property of the PACF graph, the analysis of autocorrelation coefficients estimates is advisable to perform for $\tau\in\left(0;\left[n/2\right]+1\right)$.

According to [1], $E\left(r_{PACF\ x}(\tau)\right)=-1/(n-1)$. The upper bound estimate of variance of $r_{PACF\ x}(\tau)$ calculated from (3) for arbitrary independent variables is: $Var\left(r_{PACF\ x}(\tau)\right)\le\frac{n^2}{(n-1)^2(n-2)}$. However, $Var_{norm}\left(r_{PACF\ x}(\tau)\right)=\frac{n(n-3)}{(n+1)(n-1)^2}$ [18] for the normal distribution of the initial values, and $Var_{unif}\left(r_{PACF\ x}(\tau)\right)=\frac{24}{5}n^{-2}+O\left(n^{-7/3}\log(n)\right)$ [1] for their uniform distribution. In addition, as shown in [19], the value $r_{PACF\ x}(\tau)$ is distributed asymptotically normal even for sufficiently small samples $(n>10)$. In [1] it is recommended that the estimate (5) should be between $E\left(r_{PACF\,x}(\tau)\right)-2\sqrt{Var\left(r_{PACF\,x}(\tau)\right)}$ and $E\left(r_{PACF\,x}(\tau)\right)+2\sqrt{Var\left(r_{PACF\,x}(\tau)\right)}$ for uncorrelated values $(x_0, x_1, \cdots, x_{n-1})$.

3.2 ACF Estimate on Fixed Size Sample

ACF estimate on some fixed size sample $(x_0, x_1, \cdots, x_{n-1})$ is widely used in econometrics for constructing regression models [14, 15]. In addition, a similar ACF estimate is used to investigate sequence properties in the aperiodic mode typical to information transmission in communication systems [20]. In this case, the estimate of normalized ACF is calculated according to the expression [21]:

$$r_x^*(\tau) = 1\big/(n-\tau)\cdot\sum\nolimits_{i=0}^{n-1-\tau}\left[(x_i - E(X))\cdot(x_{i+\tau}-E(X))\right]\Big/\left(1\big/n\cdot\sum\nolimits_{i=0}^{n-1}(x_i - E(X))^2\right) \quad (4)$$

for a known a priori value of $E(X)$, or the expression [21]:

$$r_x^{**}(\tau) = 1\big/(n-\tau)\cdot\sum\nolimits_{i=0}^{n-1-\tau}\left[\left(x_i-\overline{x}\right)\cdot\left(x_{i+\tau}-\overline{x}\right)\right]\Big/\left(1\big/n\cdot\sum\nolimits_{i=0}^{n-1}\left(x_i-\overline{x}\right)^2\right). \quad (5)$$

In [21] it is shown that if the random vectors $(x_i, x_{i+\tau})$ are independent and equally distributed, then the distribution law of the value $r_x^*(\tau)$ (as well as $r_x^{**}(\tau)$) has an asymptotic normal distribution.

Various authors use or recommend to use for (5) the value of zero as an approximate estimate of its expectation and the value of $n^{1/2}$ [14] or $\left(n\big/\{(n+2)(n-\tau)\}\right)^{-1/2}$ [15] as an approximate estimate of its root-mean-square deviation. However, the exact value of the $r_x^{**}(\tau)$ expectation is defined in [22] and is equal to $E\left(r_x^{**}(\tau)\right) = -(n-1)^{-1}$. The upper bound of the estimate (5) variance is defined in [23] for any law of distribution of the initial values $\{x_i\}$:

$$Var\left(r_x^{**}(\tau)\right) \le \frac{n^4-(\tau+7)n^3+(7\tau+16)n^2+2\left(\tau^2-9\tau-6\right)n-4\tau(\tau-4)}{n(n-1)^2(n-2)(n-3)}.$$

In addition, in [23] it is also shown that $Var_{norm}\left(r_x^{**}(\tau)\right) = \left(n^4-(\tau+3)n^3+3\tau n^2+2\tau(\tau+1)n-4\tau^2\right)\Big/\left((n+1)n^2(n-1)^2\right)$ for normally distributed r.v., and the expression $Var\left(r_x^{**}(\tau)\right) = (n-\tau)/\left(n(n+2)\right)$ from [15] is valid in the case of a known expectation $E(X)$, that is, for the value $r_x^*(\tau)$.

Consider the normalized ACF estimate, which is given, for example, in [9] or [10]:

$$r_x^{***}(\tau) = \sum_{i=0}^{n-1-\tau}\left[\left(x_i - \overline{x}\right)\cdot\left(x_{i+\tau} - \overline{x}\right)\right] \Big/ \sqrt{\sum_{i=0}^{n-1-\tau}\left(x_i - \overline{x}\right)^2 \cdot \sum_{i=0}^{n-1-\tau}\left(x_{i+\tau} - \overline{x}\right)^2}\ . \qquad (6)$$

Let us extend and present more fully the results of estimates of expectation $E\left(r_x^{***}(\tau)\right)$ and variance $Var\left(r_x^{***}(\tau)\right)$ obtained in [24].

The $r_x^{***}(\tau)$ distribution law for independent and equally distributed random vectors $\left(x_i, x_{i+\tau}\right)$ is an asymptotic normal distribution [25]. This statement is also confirmed by [26, 27]. They show that the distributions of correlation analysis statistics are resistant to deviations of the observed multidimensional law from normal. The empirical distributions of these statistics are well described by the boundary laws obtained from the assumption of normality of the observed values.

To find $E\left(r_x^{***}(\tau)\right)$, we use the methodology described in [22].

Let $z_i = x_i - \overline{x}$. Then $\sum_{i=0}^{n-1} z_i = \sum_{i=0}^{n-1}\left(x_i - \overline{x}\right) = \sum_{i=0}^{n-1} x_i - n\overline{x} = 0$.

$$E\left(r_x^{***}(\tau)\right) = E\left(\frac{\sum_{i=0}^{n-1-\tau} z_i \cdot z_{i+\tau}}{\sqrt{\sum_{i=0}^{n-1-\tau} z_i^2 \cdot \sum_{i=0}^{n-1-\tau} z_{i+\tau}^2}}\right) = E\left(\frac{(n-\tau)\cdot z_i \cdot z_{i+\tau}}{\sum_{i=0}^{n-1-\tau} z_i^2}\right) = n \cdot E\left(\frac{z_i \cdot z_{i+\tau}}{\sum_{i=0}^{n-1} z_i^2}\right) =$$

$$= \frac{1}{n-1}\cdot E\left(\frac{\sum_{i\neq j}\left(z_i \cdot z_j\right)}{\sum_{i=0}^{n-1} z_i^2}\right) = \frac{1}{n-1}\cdot E\left(\frac{\left(\sum_{i=0}^{n-1} z_i\right)^2 - \sum_{i=0}^{n-1} z_i^2}{\sum_{i=0}^{n-1} z_i^2}\right) = -(n-1)^{-1}.$$

The variance $Var\left(r_x^{***}(\tau)\right)$ can be calculated according to the well-known expression: $Var\left(r_x^{***}(\tau)\right) = E\left(\left(r_x^{***}(\tau)\right)^2\right) - E^2\left(r_x^{***}(\tau)\right)$.

As it shown in [23], $\left(\sum_{i=0}^{n-1-\tau} z_i z_{i+\tau}\right)^2 = \sum_{i=0}^{n-1-\tau} z_i^2 z_{i+\tau}^2 + 2\sum_{i=0}^{n-1-2\tau} z_i\, z_{i+\tau}^2 z_{i+2\tau} +$ $+\sum_{*} z_i z_{i+\tau} z_j z_{j+\tau}$. $\sum_{*}$ means a summation for $i, j = 1,\ldots,n-\tau$, and $i, i+\tau, j, j+\tau$ are different. In addition, for the symmetric joint distribution of values $z_0,\ldots,z_{n-1}$,

$$E\left(\left(r_x^{***}(\tau)\right)^2\right) = E\left[\left(\sum_{i=0}^{n-1-\tau} z_i^2 \cdot \sum_{i=0}^{n-1-\tau} z_{i+\tau}^2\right)^{-1}\begin{Bmatrix}(n-\tau)z_1^2 z_2^2 + 2(n-\tau)z_1^2 z_2 z_3 + \\ +\left((n-\tau)^2 - 2(n-2\tau) - (n-\tau)\right)z_1 z_2 z_3 z_4\end{Bmatrix}\right] =$$

$$= E\left[\left(\frac{n-\tau}{n}\sum_{i=0}^{n-1} z_i^2\right)^{-2}\begin{Bmatrix}\dfrac{n-\tau}{n(n-1)}\sum^{*} z_i^2 z_j^2 + \dfrac{2(n-2\tau)}{n(n-1)(n-2)}\sum^{*} z_i^2 z_j z_k + \\ +\dfrac{(n-\tau)^2 - 2(n-2\tau) - (n-\tau)}{n(n-1)(n-2)(n-3)}\sum^{*} z_i z_j z_k z_l\end{Bmatrix}\right],$$

where $\sum^{*}$ denotes a summation for all different indices from 0 to $n-1$.

Using an equality $S_r=\sum_{i=0}^{n-1} z_i^r$ for $r\geq 1$, and $\sum^{*} z_i^2 z_j^2 = S_2^2 - S_4$, $\sum^{*} z_i^2 z_j z_k = 2S_4 - S_2^2$, and $\sum^{*} z_i z_j z_k z_l = 3S_2^2 - 6S_4$ from [28], by analogy with [23],

$$E\left(\left(r_x^{***}(\tau)\right)^2\right)=\frac{n\begin{bmatrix}\left(-n^3+(\tau+3)n^2-\tau(n+6\tau)\right)M\left[S_4/S_2^2\right]+\\ +\left(n^2(n-\tau-4)+3(n-\tau)+3\tau(n+\tau)\right)\end{bmatrix}}{(n-1)(n-2)(n-3)(n-\tau)^2}.$$

As shown in [23, 29], $1/n \leq S_4/S_2^2 \leq 1$ for any distribution law of r.v. z_i. Then, by analogy with [23], it is possible to obtain the variance upper bound of the correlation coefficient estimate (6). To do this, replacing $E\left[S_4/S_2^2\right]$ by $1/n$ we get $E\left(\left(r_x^{***}(\tau)\right)^2\right)\leq \dfrac{n^4-n^3(\tau+5)+n^2(4\tau+6)+n\left(3\tau^2-4\tau\right)-6\tau^2}{(n-1)(n-2)(n-3)(n-\tau)^2}$. Then

$$Var\left(r_x^{***}(\tau)\right)\leq \frac{n^5-n^4(\tau+7)+n^3(7\tau+16)+n^2\left(2\tau^2-18\tau-12\right)-n\left(4\tau^2-16\tau\right)}{(n-1)^2(n-2)(n-3)(n-\tau)^2}. \quad (7)$$

According to [22], for normally distributed r.v. x_i, $E\left[S_4/S_2^2\right]=3(n-1)/\left(n(n+1)\right)$,

$$Var_{norm}\left(r_x^{***}(\tau)\right)=\frac{n^4-(\tau+3)n^3+3\tau n^2+2\tau(\tau+1)n-4\tau^2}{(n+1)(n-1)^2(n-\tau)^2}. \quad (8)$$

One can notice that $Var_{norm}\left(r_x^{***}(\tau)\right)\xrightarrow[n\gg\tau]{n\to\infty} 1/(n-\tau)$. In addition, in the general case

$$Var\left(r_x^{***}(\tau)\right)\leq \frac{1-7/n+16/n^2-12/n^3+2\tau(\tau-1)(n-2)/\left(n^3(n-\tau)\right)}{(1-1/n)^2(1-2/n)(1-3/n)(n-\tau)}\xrightarrow[n\gg\tau]{n\to\infty}\frac{1}{n-\tau}.$$

Thus, the obtained results are in full agreement with [10]. The variance of the estimate (6) is asymptotically equal to $1/(n-\tau)$ if n is large. However, the estimates (7) and (8) are more accurate.

We denote the relative error of approximation of the autocorrelation coefficient estimate variance $Var\left(r_x^{***}(\tau)\right)$ by $\delta(n,\tau)=\left|\left(Var\left(r_x^{***}(\tau)\right)-1/(n-\tau)\right)/Var\left(r_x^{***}(\tau)\right)\right|$. Analysis of the dependency $\delta(n,\tau)$ graph for $n\in[50;100]$, $\tau\in[1;25]$ (Figure 1) indicates that $\delta(n,\tau)\to\max$ for $\tau\to\max$, $n\to\min$. In particular, $\delta(100,1)=1.02\cdot 10^{-4}$, $\delta(100,25)=1.578\cdot 10^{-3}$, and $\delta(50,25)=0.02$. It is also worth noting that $\min\left(\delta(n,\tau)\right)\neq\delta(n,1)$ for a fixed value of n.

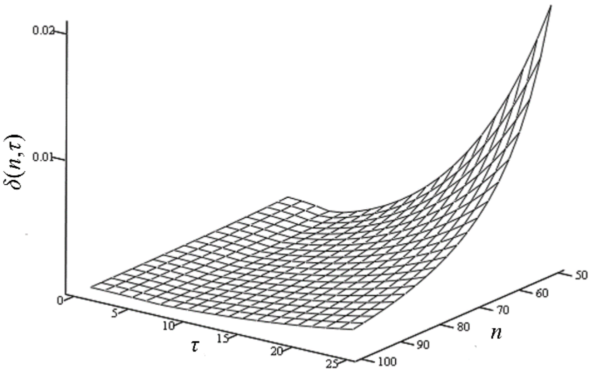

Fig. 1. The graph of dependence $\delta(n,\tau)$ for $n \in [50;100]$, $\tau \in [1;25]$

The graph of the relative error of approximation of the autocorrelation coefficient estimate variance $Var_{norm}\left(r_x^{***}(\tau)\right)$ for normally distributed r.v. x_i $\delta_{norm}(n,\tau) = \left|\left(Var_{norm}\left(r_x^{***}(\tau)\right) - 1/(n-\tau)\right) / Var_{norm}\left(r_x^{***}(\tau)\right)\right|$ depending on $n \in [50;100]$, $\tau \in [1;25]$ is shown in Figure 2.

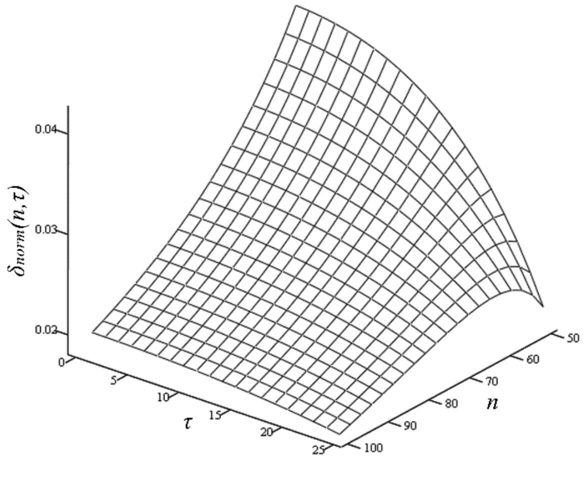

Fig. 2. The graph of dependence $\delta_{norm}(n,\tau)$ for $n \in [50;100]$, $\tau \in [1;25]$

For the specified definition area $\delta_{norm}(n,\tau) \to \max$ for $\tau \to \min$, $n \to \min$. In particular, $\delta_{norm}(100,1) = 0.021$, $\delta_{norm}(100,25) = 0.019$, and $\delta_{norm}(50,1) = 0.042$. However, in the general case $\max\left(\delta_{norm}(n,\tau)\right) \neq \delta_{norm}(n_{\min},\tau)$ for a fixed value of τ. In Figure 2, this situation is observed for $\tau \in [21;25]$.

3.3 ACF Estimate for Long Period Sequences

When analyzing the correlation properties of a random number sequence, any shift of the analyzed sequence belongs to the set $\{x_i(t)\}$ of realizations of a stationary discrete random process $X(t)$. In other words, the "zero" point for the beginning of realization may be arbitrary. In this case, we can consider a set of realizations $\{x_i(t): x_i(t) = x_{i-j}(t+j)\}$. To enable autocorrelation coefficients practical analysis, we will assume sequences of n numbers, the first of which are formed at times $t = 0, 1, 2, \ldots$, by realizations of a random process $X(t)$. This approach is most effective when the correlation properties of a random number sequence are analyzed in real time, not by some fixed size sample. In this case, the sequence elements are written to a limited size buffer (this approach is similar to the sliding window method).

We denote the element of a random number sequence at a discrete time t by x_t. Then the estimate of the normalized autocorrelation coefficient of order τ is as follows:

$$r_x(\tau) = \frac{\sum_{i=0}^{n-1}\left[\left(x_{t+i} - \overline{x(t)}\right)\cdot\left(x_{t+\tau+i} - \overline{x(t+\tau)}\right)\right]}{\sqrt{\sum_{i=0}^{n-1}\left(x_{t+i} - \overline{x(t)}\right)^2 \cdot \sum_{i=0}^{n-1}\left(x_{t+\tau+i} - \overline{x(t+\tau)}\right)^2}}, \tag{9}$$

where $\overline{x(t)} = 1/n \cdot \sum_{i=0}^{n-1} x_{t+i}$, $\overline{x(t+\tau)} = 1/n \cdot \sum_{i=0}^{n-1} x_{t+\tau+i}$ are the average sample values of random process intersections $X(t)$, $X(t+\tau)$. Note that for a stationary random process, equality $\lim_{n\to\infty}\left(P\left(\left|\overline{x(t)} - \overline{x(t+\tau)}\right| < \varepsilon\right)\right) = 1$ is satisfied for any small $\varepsilon > 0$.

The distribution law of the estimate $r = r_x(\tau)$ (9) for independent and equally distributed random vectors $(x_{t+i}, x_{t+\tau+i})$ is asymptotically normal [25] with expectation $\rho = \rho_x(\tau)$ and variance [30]

$$Var\left(r_x(\tau)\right) = \frac{\rho^2}{4n}\left(\frac{\mu_{40}}{\mu_{20}^2} + \frac{\mu_{04}}{\mu_{02}^2} + \frac{2\mu_{22}}{\mu_{20}\mu_{02}} + \frac{4\mu_{22}}{\mu_{11}^2} - \frac{4\mu_{31}}{\mu_{11}\mu_{20}} - \frac{4\mu_{13}}{\mu_{11}\mu_{02}}\right),$$

where μ_{km} are the theoretical central moments of the order k and m: $\mu_{km} = \left(x_{t+i} - E\left(x_{t+i}\right)\right)^k \left(x_{t+\tau+i} - E\left(x_{t+\tau+i}\right)\right)^m$. According to [30], $Var\left(r_x(\tau)\right) = \left(1-\rho^2\right)^2 / n$ in the case of a normally distributed population.

However, as shown in [30], applying the Fischer logarithmic transformation [31] to the sample correlation coefficients r leads to the following conclusions. The value $z = 1/2 \ln\left((1+r)/(1-r)\right)$ should be considered normally distributed with an average

$1/2\ln\left(\left(1+\rho\right)/\left(1-\rho\right)\right)+\rho/\left(2\left(n-1\right)\right)$ and a variance $1/\left(n-3\right)$. Therefore, the value $\lambda=\sqrt{n-3}\left(\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)-\left(\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)+\frac{\rho}{2\left(n-1\right)}\right)\right)$ is normal: $\lambda \square N\left(0;1\right)$.

Thus, $\rho_X\left(\tau\right)\equiv 0$ for independent and equally distributed random vectors $\left(x_{t+i},x_{t+\tau+i}\right)$, the value $\lambda_x\left(\tau\right)=\sqrt{n-3}/2\ln\left(\left(1+r_x\left(\tau\right)\right)/\left(1-r_x\left(\tau\right)\right)\right)$ has a standard normal distribution, and $\sum_{\tau=1}^{\mathrm{T}}\left(\lambda_x\left(\tau\right)\right)^2\to\chi^2_{\mathrm{T}}$

3.4 ACF Estimate for Uniformly Distributed Random/Pseudorandom Number Sequences with Known Parameters

Often, sequence statistics are checked at the output of a generator of uniformly distributed random or pseudorandom numbers in some range $\left[a,b\right]$ (such generators are most widely used for information security tasks as key entropy generators). Then the simplest analysis of the studied sequence allows us to determine the set of d.r.v. values at the generator output. If there is a sequence of numbers $\left\{x_i\in A\right\}$ from the alphabet A, then its cardinality $N=\max\left\{x_i\right\}-\min\left\{x_i\right\}+1$, the range of d.r.v. values $X\in\left[\min\left\{x_i\right\};\max\left\{x_i\right\}\right]$, its expectation $E\left(X\right)=\left(\min\left\{x_i\right\}+\max\left\{x_i\right\}\right)/2$, and its variance $Var\left(X\right)=\left(N^2-1\right)/12$.

A similar estimate can be made if the size of the analyzed number sequence significantly exceeds the capacity of the alphabet. Otherwise there is a possibility to incorrectly define the lower or upper limit of d.r.v. values set. The probability that the minimum or the maximum value of N-symbol alphabet will not be present in a sequence of V symbols is equal to $P_{er}=2\left(\left(N-1\right)/N\right)^n$.

Thus, for the given values of alphabet capacity M and probability P_{er}, it is possible to calculate the required sample size $V\geq\log_{(N-1)/N}P_{er}/2$.

For example, for $N=256$ and $P_{er}=10^{-10}$ we get: $V\geq\log_{1-1/256}10^{-10}/2=6061$.

For known parameters (expectation $E\left(X\right)$ and variance $Var\left(X\right)$) of a random process $X\left(t\right)$, the estimate of normalized ACF can be calculated by the expression

$$r_x{}'\left(\tau\right)=1/n\cdot\sum\nolimits_{i=0}^{n-1}\left[\left(x_{t+i}-E\left(X\right)\right)\cdot\left(x_{t+\tau+i}-E\left(X\right)\right)\right]/Var\left(X\right). \tag{10}$$

In this case, for independent values x_i, as well as in accordance with the regularities stated in (1) and (2), and $n\to\infty$, $r_x{}'\left(\tau\right)\to N\left(0;1/n\right)$, $\lim\limits_{\substack{n\to\infty\\ \mathrm{T}\to\infty}}\left(\sum\nolimits_{\tau=1}^{\mathrm{T}}\left(r_x{}'\left(\tau\right)\right)^2\right)=\mathrm{T}/n$,

$$n\sum_{\tau=1}^{\mathrm{T}}\left(r_x{}'(\tau)\right)^2 \to \chi^2_{\mathrm{T}}\,. \tag{11}$$

For side lobes power $W'(\mathrm{T})=\sum_{\tau=1}^{\mathrm{T}}\left(r_x{}'(\tau)\right)^2$ of ACF estimate (10), $\lim\limits_{\substack{n\to\infty\\ \mathrm{T}\to\infty}}\left(W'(\mathrm{T})\right)=\mathrm{T}/n$ and the expression (11) can be rewritten as follows:

$$nW'(\mathrm{T})\to\chi^2_{\mathrm{T}}\,. \tag{12}$$

In this case, the statistical criterion for corresponding an ACF estimate of uniformly distributed random number sequence to the white noise ACF provides calculating the normalized autocorrelation coefficients from (10), forming an estimate of the side lobes power $W'(\mathrm{T})$, and then estimating it with (12).

4 Description of the Criterion for Estimating Uniformly Distributed Random Number Sequences

The criterion for estimating uniformly distributed random number sequences is the following:

1. if the d.r.v. definition area is unknown, it is empirically determined;
2. the d.r.v. expectation and variance are calculated;
3. using the expression (10), the sequence of estimates $r_x{}'(\tau)$ of normalized autocorrelation coefficients is calculated for $\tau\in[1;\mathrm{T}]$;
4. the side lobes power $W'(\mathrm{T})=\sum_{\tau=1}^{\mathrm{T}}\left(r_x{}'(\tau)\right)^2$ of the ACF estimate is calculated;
5. if the value $nW'(\mathrm{T})$ for the selected level of significance does not exceed the quantile $\chi^2_{1-\alpha,\mathrm{T}}$ of chi-square distribution with T degrees of freedom, the null hypothesis is accepted. It is that the numbers of the sequence under study are random. Otherwise, the null hypothesis is rejected.

5 Applying the Criterion for Estimating Uniformly Distributed Random Number Sequences

We implement the developed criterion for known PRNG, which pass all tests of the TestU01 test package [4]. For this purpose we conduct $N=1000$ independent tests. We define the value $nW'(\mathrm{T})$. Then calculate the relative frequency of the event $A=\left\{nW'(\mathrm{T})\le\chi^2_{1-\alpha,\mathrm{T}}\right\}$.

We denote by Q the value that in each specific test takes a value of 1 if the event A is true and 0 if the event A is false. Then the relative frequency of the event $A=\left\{nW'(\mathrm{T}) \leq \chi^2_{1-\alpha,\mathrm{T}}\right\}$ in N independent tests is $p^* = \sum_{i=1}^{N} Q_i / N$.

The expectation of the relative frequency is $E(p^*) = 1-\alpha$, its variance is $Var(p^*) = \alpha(1-\alpha)/N$. Then the inequality $\left|p^* - (1-\alpha)\right| < t_\gamma \sqrt{\alpha(1-\alpha)/N}$ is satisfied with probability γ, where t_γ is the quantile of standard normal distribution with level γ. In other words, the calculated relative frequency p^* falls within the confidence interval $\left(1-\alpha-t_\gamma\sqrt{\alpha(1-\alpha)/N}; 1-\alpha+t_\gamma\sqrt{\alpha(1-\alpha)/N}\right)$ with probability γ.

For testing we will choose $\alpha = \gamma = 0.05$. Then the confidence interval for p^* is $(0.9365; 0.9635)$. The test results are summarized in Table 1.

Table 1. The results of applying the criterion for estimating uniformly distributed PRNG

PRNG	TestU01 autocorrelation tests results (4 smallCrush + 4 bigCrush), passed/total tests	Test results of the developed criterion
LCG(2^{24}, 16598013, 12820163)	2/8	failed
LCG(2^{31}, 65539, 0)	6/8	failed
LCG(2^{32}, 69069, 1)	5/8	failed
LCG(2^{32}, 1099087573, 0)	5/8	failed
LCG(2^{46}, 5^{13}, 0)	7/8	failed
LCG(2^{48}, 25214903917, 11)	7/8	failed
LCG(2^{48}, 5^{19}, 0)	7/8	failed
LCG(2^{48}, 33952834046453, 0)	7/8	failed
LCG(2^{48}, 44485709377909, 0)	7/8	failed
LCG(2^{59}, 3^{13}, 0)	8/8	failed
LCG(2^{63}, 5^{19}, 1)	8/8	failed
LCG(2^{63}, 9219741426499971445, 1)	8/8	failed
LCG($2^{31}-1$, $2^{31}-2^{10}$, 0)	6/8	passed
LCG($2^{31}-1$, 16807, 0)	7/8	passed
LCG($2^{61}-1$, $2^{30}-2^{19}$, 0)	8/8	passed
LCG($10^{12}-11$,427419669081, 0)	8/8	passed

The results indicate that some of the generators that successfully pass all the TestU01 autocorrelation tests do not meet the developed criterion.

6 Conclusions

The study has produced the following results:

- integral estimate of normalized autocorrelation coefficients (ACF side lobes) is theoretically obtained. This provides the basis for building statistical criteria for verifying correlation properties of random and pseudorandom number sequences by their empirical estimates;
- the first and second initial moments of estimates of normalized autocorrelation coefficients are presented for: PACF; ACF of fixed size sample calculated according to different approaches (for example, presented in [9, 10, 21]); "sliding window" ACF for long period sequences;
- the upper bound of the variance of estimates of normalized autocorrelation coefficients calculated by [9] or [10] is clarified. This allows increasing the accuracy of ACF side lobes integral estimate;
- the criterion for estimating autocorrelation of time series based on simultaneous analysis of several autocorrelation coefficients (similar to the Box-Pierce [14] and Ljung–Box [15] criteria) has been further developed by adapting it to uniformly distributed r.v. This made it possible to perform a complex estimate of ACF for sequences of uniformly distributed random and pseudorandom numbers;
- applying the criterion revealed statistical deviations for some PRNG that successfully pass all TestU01 autocorrelation tests.

References

1. Knuth, D. E.: The Art of Computer Programming: Seminumerical Algorithms, 3 ed., vol. 2. Boston, Addison-Wesley Longman Publishing Co., Inc. (1997)
2. Marsaglia, G.: DIEHARD Battery of Tests of Randomness. http://www.stat.fsu.edu/pub/diehard
3. Bassham, L. E. et al.: A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications. SP 800-22 Rev. 1a., Gaithersburg, MD, United States (2010)
4. L'Ecuyer, P., Simard, R.: TestU01: A C library for empirical testing of random number generators. ACM Transactions on Mathematical Software. **33**, no. 4, 22-es (2007). doi: 10.1145/1268776.1268777
5. Caelli, W., Dawson, E., Nielsen, L., Gustafson, H.: CRYPT–X Statistical Package Manual, Measuring the strength of Stream and Block Ciphers. Queensland University of Technology (1992)
6. Security requirements for cryptographic modules. US standard FIPS PUB 140-2 (2001)
7. Papoulis, A., Pillai, S. U.: Probability, random variables, and stochastic processes, 4th ed. Boston, McGraw-Hill (2002)
8. Ivanov, M.A., Chugunkov, I.V.: Theory, application and quality evaluation of pseudorandom sequence generators. Moscow, Kudits-Obraz (2003) (in Russian)
9. Smirnov, N.V., Dunin-Barkovsky, I.V.: Course in probability theory and mathematical statistics for technical applications. Moscow, Nauka (1969) (in Russian)
10. Kendall, M.G.: The advanced theory of statistics, vol. 2. London: C. Griffin & Company limited (1946)
11. Faure, E.V., Shcherba, A.I., Rudnytskyi, V.M.: The method and criterion for quality assessment of random number sequences. Cybernetics and Systems Analysis, **52**, no. 2, 277–284 (2016). doi: 10.1007/s10559-016-9824-3

12. Dirac, P.A.M.: The principles of quantum mechanics. London, OUP (1958). doi: 10.1063/1.3062610
13. Bolshev, L.N., Smirnov, N.V.: Tables of mathematical statistics. Moscow, Nauka (1983) (in Russian)
14. Box, G.E.P., Pierce, D.A.: Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. Journal of the American Statistical Association, **65**, no. 332, 1509–1526 (1970). doi: 10.1080/01621459.1970.10481180
15. Ljung, G.M., Box, G.E.P.: On a measure of lack of fit in time series models. Biometrika, **65**, no. 2, 297 (1978). doi: 10.1093/biomet/65.2.297
16. Wentzel, Ye.S., Ovcharov, L.A.: Applied problems of probabilities theory. Moscow, Radio and Communication (1983) (in Russian)
17. Kuznetsov, V.M.: Generators of random and pseudorandom sequences on digital delay elements (basics of theory and methods of construction). Thesis for doctor of technical sciences degree. Kazan, Kazan State Technical University (2011) (in Russian)
18. Dixon, W.J.: Further contributions to the problem of serial correlation. Ann. Math. Statist., **15**, no. 2, 119–144 (1944). doi: 10.1214/aoms/1177731279
19. Lemeshko, B.Yu., Komissarov, A.S., Shcheglov, A.Ye.: Application of tests for trend detection and checking for randomness. Metrology, **12**, 3–25 (2010) (in Russian)
20. Varakin, L.E.: Communication systems with noise-like signals. Moscow, Radio and Communication (1985) (in Russian)
21. Anderson, T.W., Walker, A.M.: On the asymptotic distribution of the autocorrelations of a sample from a linear stochastic process. The Annals of Mathematical Statistics, **35**, no. 3, 1296–1303 (1964). doi: 10.1214/aoms/1177703285
22. Moran, P.A.P.: Some theorems on time series: II the significance of the serial correlation coefficient. Biometrika, **35**, no. 3/4, 255–260 (1948). doi: 10.1093/biomet/35.3-4.255
23. Dufour, J.-M., Roy, R.: Some robust exact results on sample autocorrelations and tests of randomness. Journal of Econometrics, **29**, no. 3, 257–273 (1985). doi: 10.1016/0304-4076(85)90155-1
24. Faure, E.V.: Statistical characteristics of estimates of normalized autocorrelation coefficients of (pseudo) random number sequences. In: All-Ukrainian Scientific and Practical Internet Conference on Automation and Computer-Integrated Technologies in Production and Education: State, Achievements, Prospects of Development, pp. 46-47. Cherkasy (2015) (in Russian)
25. Orlov, A.I.: Applied statistics. Moscow, Examen (2004) (in Russian)
26. Lemeshko, B.Yu., Pomadin, S.S.: Correlation analysis of observations of many-dimensional random variables under violation of normality assumptions. Siberian Journal of Industrial Mathematics, **5**, no. 3 (11), 115-130 (2002) (in Russian)
27. Pomadin, S.S.: Investigation of statistics distributions of multidimensional data analysis in violation of normality assumptions, dissertation. Thesis for candidate of technical sciences degree. Novosibirsk, Novosibirsk State Technical University (2004) (in Russian)
28. Kendall, M.G., Stuart, A., Ord, J.K.: The advanced theory of statistics. 4 ed. V. 3: Design and analysis, and time-series. London: C. Griffin & Company limited (1983)
29. Moran P.A.P.: Testing for serial correlation with exponentially distributed variates. Biometrika, **54**, no. 3/4, 395–401 (1967). doi: 10.1093/biomet/54.3-4.395
30. Kramer, G.: Mathematical methods of statistics, 2 ed. Moscow, Mir (1975)
31. Fisher, R.A.: On the probable error of a coefficient of correlation deduced from a small sample. Metron, **1**, 3–32 (1921)