Polish Development and Research Project within the Smart Growth Operational Programme SGOP 1.1.1.1 of The National Centre for Research and Development: JURAND - National Quantum Random Number Generator (POIR.01.01.01-00-0173/15-00)

Random numbers generator statistical tests

W. Jacak, J. Jacak, W. Donderowicz, and L. Jacak

CONTENTS

I. RANDOM BIT SEQUENCE

The basic model of a random bit sequence is a repeatable event of a measurement which gives only two possible results (labeled as 0 and 1), where each result is absolutely independent form previous results – such situation is commonly modelled as coin flipping with use of an unbiased coin (50% of heads and 50% of tails). In a random bit sequence each bit is generated independently of previous bits, which means that regardless of how many elements of such sequence are already know, the resultant of the next bit cannot be predicted.

Above model is an idealization of a random bit sequence generator which cannot be implemented with use of classical information science due to the deterministic nature of classical physics – each generator based solely on classical physics mechanisms will always work according to some deterministic process (even highly complex but still predictable). Thus it is a common fact that classical random number generators cannot produce a real random bit sequence.

But in the case of a quantum methods of information processing or generally quantum physics such ideal model is achivable. Quantum mechanics provides some fundamental rules which allows to construct a simple system which acording to a so called von Neumann measurement scheme [1] will allow to generate independently 2 possible values with probability equalt to 50%. Thus this is why the quantum random number generators are of such interest for morder applications in information security area.

A. Von Neumann's scheme

B. Testing the randomness of a bit sequence

One of the most important aspect of generating a random sequence of bits is testing whether it is random or not. It is stated that randomness is a probabilistic property, which allows to charactirize a random bit sequence in terms of probability.

Each seqence can be analized in comparison to truly random seqence expressed in probabilitic values.

But there is a fundamental problem – there exists an infinit number of possible statistical test, each corresponding to some unique pattern. Each test assess whether such pattern is present in tested sequence or not (if the patter is present then such sequence is considered as nonrandom). Thus it is rather impossible to find a complete set of test which verify if a sequence is truly random.

II. NIST TESTS OVERVIEW

A. Frequency test

Based on [2–4].

In a truly random sequence the number of occurrences in the entire sequence should be the same for zeros and ones. One would expect that the fraction of the total number of zeros over the total lenght of the sequence or the total number of ones over the total lenght length of the sequence should be close to $\frac{1}{2}$. This test bases on sigle bits of sequence.

Identically distributed Bernoulli random varibale corresponds to binary value with probability equal $\frac{1}{2}$ of getting 1 or 0 value.

The de Moivre-Laplace theorem states that for a sufficiently large number of trials the distribution of a sum of Bernoulli random variables normalized by $\sqrt{(n)}$ can be approxiamted by a standard normal distribution (in other words binomial distribution with some additionals conditions can be approximated by a standard normal distribution).

Here one can use a approxiamtion assessing the closeness of the fraction of ones to $\frac{1}{2}$.

Using the Central Limit Theorem which states that, for certain conditions, the sum of sufficiently large numbers of iterates of Bernoulli random variables can be approximated by normal distribution [5], one can write a limit

$$\lim_{n\to\infty} P\left(\frac{\sum_{i=1}^{n}(2\varepsilon_i-1)}{\sqrt{n}}\le z\right)=\Phi(z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}e^{-\frac{u^2}{2}}du, \tag{1}$$

where $\Phi(z)$ is a standard normal cumulative distribution function, and

$$\begin{aligned}\Phi(z)&=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}e^{-\frac{u^2}{2}}du=\frac{1}{\sqrt{\pi}}\int_{-\infty}^{\frac{z}{\sqrt{2}}}e^{-u^2}du=\frac{1}{\sqrt{\pi}}\int_{0}^{\frac{z}{\sqrt{2}}}e^{-u^2}du+\frac{1}{\sqrt{\pi}}\int_{-\infty}^{0}e^{-u^2}du\\&=\frac{1}{2}\left(\frac{2}{\sqrt{\pi}}\int_{0}^{\frac{z}{\sqrt{2}}}e^{-u^2}du+\frac{2}{\sqrt{\pi}}\int_{-\infty}^{0}e^{-u^2}du\right)=\frac{1}{2}\left(\operatorname{erf}\left(\frac{z}{\sqrt{2}}\right)+1\right),\end{aligned} \tag{2}$$

where error function $\operatorname{erf}(z)=\frac{2}{\sqrt{\pi}}\int_0^z e^{-u^2}du$. In this test a positive z is assumed, as there is no difference if the number of ones or the number of zeros exceed the other one in a non random manner. One can write

$$\begin{aligned}&\Phi(z)=1-\Phi(-z),\\&\Phi(-z,z)=\Phi(z)-\Phi(-z)=\Phi(z)-(1-\Phi(z))=2\Phi(z)-1,\\&\Phi(-z,z)^c=1-\Phi(-z,z)=2(1-\Phi(z)),\\&2(1-\Phi(z))=2\left(1-\frac{1}{2}\left(\operatorname{erf}\left(\frac{z}{\sqrt{2}}\right)+1\right)\right)=1-\operatorname{erf}\left(\frac{z}{\sqrt{2}}\right)=\operatorname{erfc}\left(\frac{z}{\sqrt{2}}\right)\end{aligned} \tag{3}$$

thus for the statistics $S_{\text{obs}}=\frac{|S_n|}{\sqrt{n}}, S_n=\sum_{i=1}^{n}(2\varepsilon_i-1)$ the P-value has to form

$$P\text{-value}=\operatorname{erfc}\left(\frac{S_{\text{obs}}}{\sqrt{2}}\right), \tag{4}$$

where complementary error function has the form $\operatorname{erfc}(z)=\frac{2}{\sqrt{\pi}}\int_z^\infty e^{u^2}du$.

The test can be perfomed in the following manner

1. Convert all zeros to -1 in the sequnce, then add all values, $S_n = \sum_{i=1}^{n} (2\varepsilon_i - 1)$, where ε_i is the value of the ith possition in the sequence.
2. Compute the statistics $S_{\text{obs}} = \frac{|S_n|}{\sqrt{n}}$.
3. Compute the P-value given by the complementary error function, $P\text{-value} = \text{erfc}\left(\frac{S_{\text{obs}}}{\sqrt{2}}\right)$.
4. If the P-value is greater than assumed significance level (typically choosen within the range from 0.001 to 0.01) then the sequence can be considered random.

This test is considered as fundamental, as other tests base on the result of this one. It is worth to notice that this test will be passed by a sequence where first half is ones and second is zeros – which is obviously not a random sequence – thus a test which analyzies the proper distribution of ones and zeros in whole sequence is needed.

If the obtained P-value is lower then accepted threshold (0.01) then the sequence is non-random.

It is recommended that tested sequences consist of a mimimum of 100 bits.

B. Frequency test within a block

Based on [2, 6–8].

This test analizes the proportion of the number of occurrences of zeros and ones in each of N blocks of bit length M, where $MN = n$ is the number of bits in testes bit sequence. It checks whether the deviations from the ideal proportion, $\frac{1}{2}$, of ones in each block, can be considered as random or not. In case of $M = 1$ this test is the same as frequency test.

A chi-square test is applied to the initial sequence divided on N nonoverlapping blocks of length M, comparing calculated proportion of ones in substring to $\frac{1}{2}$:

$$\chi^2_{\text{obs}} = 4M \sum_{i=1}^{N} \left(\pi_i - \frac{1}{2}\right)^2, \tag{5}$$

where $\pi_i = \frac{\sum_{k=1}^{M} \varepsilon_k^{(i)}}{M}$, and $\varepsilon_k^{(i)}$ it the value of the kth bit in the ith block.

Next the P-value is calculated with use of the incomplete gamma function for $Q(a, x)$ defined as follows

$$Q(a, x) = 1 - P(a, x) = \frac{\Gamma(a, x)}{\Gamma(a)} = \frac{1}{\Gamma(a)} \int_x^{\infty} e^{-t} t^{a-1} dt, \tag{6}$$

where $Q(a, 0) = 1$, $Q(a, \infty) = 0$, and $\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$.

The P-value has the form

$$\frac{\int_{\chi^2_{\text{obs}}}^{\infty} e^{-\frac{u}{2}} u^{\frac{N}{2}-1} du}{\Gamma\left(\frac{N}{2}\right) 2^{\frac{N}{2}}} = \frac{\int_{\frac{\chi^2_{\text{obs}}}{2}}^{\infty} e^{-u} u^{\frac{N}{2}-1} du}{\Gamma\left(\frac{N}{2}\right)} \tag{7}$$

which can be written as so called complementary regularized upper incomplete gamma function, $\Gamma_C(s, x) = \frac{1}{\Gamma(s)} \int_x^{\infty} t^{s-1} e^t dt$, giving

$$P\text{-value} = \text{igamc}\left(\frac{N}{2}, \frac{\chi^2_{\text{obs}}}{2}\right). \tag{8}$$

If the obtained P-value is lower then accepted threshold (0.01) then the sequence is non-random.

It is recommended that tested sequences consist of a mimimum of 100 bits. The block size M should be selected such that $M \geq 20, M > 0.01n$ and $N < 100$

C. Runs test

Based on [2, 9, 10].

A run of lenght k is defined as a sequence of consecutive identical bits of length k. Such sequence is bounded befor and after with bits of the opposite value. Runs test analizes the total number of runs in the sequence determining

if the number of runs of zeros and ones with various lenghts stays similar to numbers for a truly random sequence. This test can give some information about the transitions between zeros and ones, whether they happen too often or too rarely.

To calculate te number of runs a function $r(k)$ is deifiened aspect

$$r(k) = \begin{cases} 0, & \text{for} \quad \varepsilon_k = \varepsilon_{k+1}, \\ 1, & \text{for} \quad \varepsilon_k \neq \varepsilon_{k+1}, \end{cases} \quad k = 1, \ldots, n-1. \tag{9}$$

The total number of runs in the n-bit sequence, $\varepsilon = \varepsilon_1 \varepsilon_2 \ldots \varepsilon_n$, defined as V_n can be calculated as follows

$$V_n = \sum_{k=1}^{n-1} r(k) + 1. \tag{10}$$

Its assumed that the distribution of the total number of runs, V_n is tends to normal with growing n, for fixed proportion $\pi = \sum_j \frac{\varepsilon_j}{n}$ (which is checked in frequency test bo be close to $\frac{1}{2}$, $\left|\pi - \frac{1}{2}\right| \leq \frac{2}{\sqrt{n}}$, otherwise the P-value is set to 0)

$$\lim_{n\to\infty} P\left(\frac{V_N - 2n\pi(1-\pi)}{2\sqrt{n}\pi(1-\pi)} \leq z\right) = \Phi(z). \tag{11}$$

Thus the P-value is calculated as the complementary error function

$$P\text{-value} = \operatorname{erfc}\left(\frac{|V_n(\text{obs}) - 2n\pi(1-\pi)|}{2\sqrt{2n}\pi(1-\pi)}\right). \tag{12}$$

If the obtained P-value is lower then accepted threshold (0.01) then the sequence is non-random.

It is recommended that tested sequences consist of a mimimum of 100 bits.

D. Longest run of ones in a block test

Based on [2, 10–12].

Another factor for characterisation of randomness is the longest consecutive subsequence of ones. This test operates on N blocks of length M, where NM is the lenght of tested bit sequence. Depending of the overal length, n, of the analized sequence, one sould choose apropriate M, eg. $n = 128 \to M = 8$, $n = 6272 \to M = 128$, $n = 750000 \to M = 10^4$. Basing on choosed M other parameters can be selected, $M = 8 \to K = 3, N = 16$, $M = 128 \to K = 5, N = 49$, $M = 10^4 \to K = 6, N = 75$. The value $K + 1$ defines the number of classes of runs of ones according to table I. For each such class one should calculte how many runs (of length according to tabel I) are present in analized block of lenght M.

K	M	Class	Probability
3	8	$\{\nu \leq 1\}$	$\pi_0 = 0.2148$
3	8	$\{\nu = 2\}$	$\pi_1 = 0.3672$
3	8	$\{\nu = 3\}$	$\pi_2 = 0.2305$
3	8	$\{\nu \geq 4\}$	$\pi_3 = 0.1875$

K	M	Class	Probability
5	128	$\{\nu \leq 4\}$	$\pi_0 = 0.1174$
5	128	$\{\nu = 5\}$	$\pi_1 = 0.2430$
5	128	$\{\nu = 6\}$	$\pi_2 = 0.2493$
5	128	$\{\nu = 7\}$	$\pi_3 = 0.1752$
5	128	$\{\nu = 8\}$	$\pi_4 = 0.1027$
5	128	$\{\nu \geq 9\}$	$\pi_5 = 0.1124$

K	M	Class	Probability
5	512	$\{\nu \leq 6\}$	$\pi_0 = 0.1170$
5	512	$\{\nu = 7\}$	$\pi_1 = 0.2460$
5	512	$\{\nu = 8\}$	$\pi_2 = 0.2523$
5	512	$\{\nu = 9\}$	$\pi_3 = 0.1755$
5	512	$\{\nu = 10\}$	$\pi_4 = 0.1027$
5	512	$\{\nu \geq 11\}$	$\pi_5 = 0.1124$

K	M	Class	Probability
5	1000	$\{\nu \leq 7\}$	$\pi_0 = 0.1307$
5	1000	$\{\nu = 8\}$	$\pi_1 = 0.2437$
5	1000	$\{\nu = 9\}$	$\pi_2 = 0.2452$
5	1000	$\{\nu = 10\}$	$\pi_3 = 0.1714$
5	1000	$\{\nu = 11\}$	$\pi_4 = 0.1002$
5	1000	$\{\nu \geq 12\}$	$\pi_5 = 0.1088$

K	M	Class	Probability
6	10000	$\{\nu \leq 10\}$	$\pi_0 = 0.0882$
6	10000	$\{\nu = 11\}$	$\pi_1 = 0.2092$
6	10000	$\{\nu = 12\}$	$\pi_2 = 0.2483$
6	10000	$\{\nu = 13\}$	$\pi_3 = 0.1933$
6	10000	$\{\nu = 14\}$	$\pi_4 = 0.1208$
6	10000	$\{\nu \geq 15\}$	$\pi_5 = 0.0675$
6	10000	$\{\nu \geq 16\}$	$\pi_5 = 0.0727$

TABLE I: Classes and probabilities for different K and M

To obtain the empirical frequencies needed to define the P-value, the conditional probability for the longest run of ones ν should calculated as follows

$$P\left(\nu \leq m|r\right) = \frac{1}{\binom{M}{r}} \sum_{j=0}^{U} (-1)^j \binom{M-r+1}{j} \binom{M-j\left(m+1\right)}{M-r}, \tag{13}$$

where r is the number of ones, $M-r$ the number of zeros in the m-bit block, $U = \min\left(M-r+1, \left[\frac{r}{m+1}\right]\right)$. Thus

$$P\left(\nu \leq m\right) = \sum_{r=0}^{M} \binom{M}{r} P\left(\nu \leq m|r\right) \frac{1}{2^M}. \tag{14}$$

The theoretical probabilites, $\pi_0, \pi_1, \ldots, \pi_K$ (presented in table I), and the empirical frequencies, $\nu_0, \nu_1, \ldots, \nu_K$ are conjoined by a chi-square defined as follows

$$\chi^2 = \sum_{i=0}^{K} \frac{\left(\nu_i - N\pi_i\right)^2}{N\pi_i}, \tag{15}$$

which for random sequnce has an approximate chi-square distribution with K degrees of freedom.

The analized P-value has the form

$$\frac{\int_{\chi^2_{\text{obs}}}^{\infty} e^{-\frac{u}{2}} u^{\frac{K}{2}-1} du}{\Gamma\left(\frac{K}{2}\right) 2^{\frac{K}{2}}} = \text{igamc}\left(\frac{K}{2}, \frac{\chi^2_{\text{obs}}}{2}\right), \tag{16}$$

where igamc is the complementary regularized upper incomplete gamma function.

If the obtained P-value is lower then accepted threshold (0.01) then the sequence is non-random.

It is recommended that minimal number of bits in tested sequence should be 128, 6272 or 750000.

E. Binary matrix rank test

Based on [2, 13–15].

This test is also a part of DIEHARD tests.

It is possible to check the randomness by testing the linear dependences among same length substrings of tested sequence. One can construct matrices form tested seqence (by dividing it on consequtive subsequences as rows or columns) and find linear correlations between rows or columns in such matrices. In such matrices the deviation of the rank (or its deficiency) from theoretically obtain values can be tested form the statistical point of view.

In case of random binary matrices of dimension $M \times Q$ theirs rank can attain the value $r = 0, 1, 2, \ldots, m$, where $m = \min M, Q$, with probabilites defined as

$$p_r = 2^{r(Q+M-r)-MQ} \prod_{i=0}^{r-1} \frac{\left(1-2^{i-Q}\right)\left(1-2^{i-M}\right)}{1-2^{i-r}}. \tag{17}$$

Values of M and Q can be choosen as equal, eg. $M = Q = 32$, and then the number M will be only parameter of this test. In case when $n = M^2N$, then N will be new sample size, and when $n \neq M^2N$ then M and N are choosen so that $n - M^2N$ which will be discarded is relatively/negligibly small.

In practice the $N = \left[\frac{n}{MQ}\right]$ discarding some small number of bits in each subsequence.

It can be justified based on probabilities in the case of theoretical random sequence for $M = Q = 32$, $p_M \approx \prod_{j=1}^{\infty}\left[1-\frac{1}{2^j}\right] = 0.2888\ldots$, $p_{M-1} \approx 2p_M \approx 0.5776\ldots$, $p_{M-2} \approx \frac{4}{9}p_M \approx 0.1284\ldots$ – all other probabilites, for $M \geq 10$ are relatively small, ≤ 0.005.

If one has N square matrices of size $M \times M$ then for each matrix the rank is evaluated according to below scheme:

After ranks evaluation, the frequencies F_M – number of matrices with rank equal M, F_{M-1} – number of matrices with rank equal to $M-1$ and $N - F_M - F_{M-1}$ – number of matrices with rank not exceeding $M-2$, are calculated.

The statistics is defined as follows

$$\chi^2 = \frac{\left(F_M - p_M N\right)^2}{p_M N} + \frac{\left(F_{M-1} - p_{M-1} N\right)^2}{p_{M-1} N} + \frac{\left(N - F_M - F_{M-1} - \left(1 - p_M - p_{M-1}\right) N\right)^2}{\left(1 - p_M - p_{M-1}\right) N}, \tag{18}$$

which under assumption of randomness of tested sequence should have an approximate chi-square distribution with 2 degrees of freedom.

Analized P-value is

$$P\text{-value} = e^{-\frac{\chi^2_{\text{obs}}}{2}}. \tag{19}$$

If χ^2_{obs} is large then the deviation of rank distribution from distribution for random sequnce is significant.

If the obtained P-value is lower then accepted threshold (0.01) then the sequence is non-random.

It is recommended to select n such that in case of $M = Q = 32$, $n \geq 38MQ$, thus $n = 38912$.

F. Discrete Fourier transform (spectral) test

Based on [2, 16–18].

This test is one of the spectral methods based on discrete Fourier transfrom. It finds the periodic behaviour of analized bit sequence, which would indicate the lack of randomness.

In tested sequnce bits are coded -1 and $+1$. Discrete Fourier transform has the form

$$f_j = \sum_{k=1}^{n} x_k e^{i2\pi \frac{(k-1)j}{n}}, \tag{20}$$

where x_k corresponds to bit on position k in initial sequence, $k = 1, \ldots, n$, $j = 0, \ldots, n$. As the Fourier transform is symmetrical in real and complex part, thus only values from 0 to $\frac{n}{2} - 1$ are considered here. In case of randomness of x_k series the modulus of Fourier transform, $|f_j|$ should be smaller then $h = \sqrt{\left(\log \frac{1}{0.05}\right) n}$ in 95% cases. P-value here is derived from binomial distribution.

$$d = \frac{N_1 - N_0}{\sqrt{\frac{n(0.95)(0.05)}{4}}}, \tag{21}$$

where N_1 is the number of peaks less then h – only first half of peaks are considered ($\frac{n}{2}$ – due to symmetry of real and complex part of transform). The P-value is defined as follows

$$2\left(1 - \phi\left(|d|\right)\right) = \operatorname{erfc}\left(\frac{|d|}{\sqrt{2}}\right), \tag{22}$$

where $\phi(x)$ is the cumulative probability function of the standard normal distribution, and erfc is a complementary error function.

It is possible to define also other P-values on Fourier transfrom or its modulus to analyzie the deviation form randomness behaviour of bit sequnece.

If the obtained P-value is lower then accepted threshold (0.01) then the sequence is non-random.

It is recommended to apply this test to a sequence of length at least 1000 bits.

G. Non-overlapping template matching test

Based on [2, 19].

This test analyzes occurrences of some pre-defined patterns, of a non-periodic character, in order to detect whether if there are too many of theirs occurances. Analysing the input sequence with a windows of size m configured to detect specific pattern of size m, in case when patter was not found the windows moves one bit and checks again, in case when patter was found the window moves to bit after the last bit of found patter to search for next occurences.

Patterns can be defined as follows

$$B = \left(\varepsilon_1^0, \ldots, \varepsilon_m^0\right), \tag{23}$$

where m is fixed length. Patterns will be choosen as parameters for this test. Exemplary aperiodic patterns are presented in table II.

Pattern B can be characterized by its set of periods B

$$\mathrm{B} = \left\{j, 1 \leq j \leq m-1, \varepsilon_{j+k}^0 = \varepsilon_k^0, k = 1, \ldots, m-j\right\}, \tag{24}$$

for example, for B corresponding to a string of only ones of length m then its set of periods B $= \{1, \ldots, m-1\}$. In case of aperiodic patterns B the set of periods B is empty. Such pattern cannot be written in form of $ll \ldots ll^{'}$ for l shorter than B with $l^{'}$ denoting a prefix of l. In such cases occurrences of B in the analyzed bit sequence are non-overlapping.

In considered test the number of occurences of the given patterns is analized which plays the role of statistics. It is defined as

$$W = W(m,n) = \sum_{i=1}^{n-m+1} I\left(\varepsilon_{i+k-1} = \varepsilon_k^0, k = 1, ..., m\right), \tag{25}$$

where m is the pattern B length, ε_i is the i-th bit of analyzed sequence (of length M) and ε_i^0 is the i-th bit of given pattern. The statistic W is defined also for aperiodic patterns, where B $= \emptyset$. In considered case the Central Limit Theorem holds for statistics W as the random variables $I\left(\varepsilon_{i+k-1} = \varepsilon_k^0, k = 1, ..., m\right)$ are dependent of m. Normal distribution approximation parameters, mean and variance, have the form

$$\mu = \frac{n-m+1}{2^m}, \quad \sigma^2 = n\left(\frac{1}{2^m} - \frac{2m-1}{2^{2m}}\right), \tag{26}$$

where $n = MN$ is the total lenght of analyzed sequence, N is the number of blocks of length M.

$m=2$	$m=3$	$m=4$	$m=5$	$m=6$	$m=7$	$m=8$
01	001	0001	00001	000001	0000001	00000001
10	011	0011	00011	000011	0000011	00000011
	100	0111	00101	000101	0000101	00000101
	110	1000	01011	000111	0000111	00000111
		1100	00111	001011	0001001	00001001
		1110	01111	001101	0001011	00001011
			11100	001111	0001101	00001101
			11010	010011	0001111	00001111
			10100	010111	0010011	00010011
			11000	011111	0010101	00010101
			10000	100000	0010111	00010111
			11110	101000	0011011	00011001
				101100	0011101	00011011
				110000	0011111	00011101
				110010	0100011	00011111
				110100	0100111	00100011
				111000	0101011	00100101
				111010	0101111	00100111
				111100	0110111	00101011
				111110	0111111	00101101
					1000000	00101111
					1001000	00110101
					1010000	00110111
					1010100	00111011
					1011000	00111101
					1011100	00111111
					1100000	01000011
					1100010	01000111
					1100100	01001011
					1101000	01001111
					1101010	01010011
					1101100	01010111
					1110000	01011011
					1110010	01011111
					1110100	01100111
					1110110	01101111
					1111000	01111111

$m=2$	$m=3$	$m=4$	$m=5$	$m=6$	$m=7$	$m=8$
					1111010	10000000
					1111100	10010000
					1111110	10011000
						10100000
						10100100
						10101000
						10101100
						10110000
						10110100
						10111000
						10111100
						11000000
						11000010
						11000100
						11001000
						11001010
						11010000
						11010010
						11010100
						11011000
						11011010
						11011100
						11100000
						11100010
						11100100
						11100110
						11101000
						11101010
						11101100
						11110000
						11110010
						11110100
						11110110
						11111000
						11111010
						11111100
						11111110

TABLE II: Aperiodic patterns for small m

For each j-th block ($j = 1, \ldots, N$) the statistics of the number of occurrences of pattern B is calculated, $W_j = W_j(m, M)$. For each W_j, let $\mu = (M-m+1)\,2^{-m}$ and $\sigma^2 = M\left(\frac{1}{2^m} - \frac{2m-1}{2^{2m}}\right)$. If M is sufficiently large then W_j has a normal distribution with mean μ and variance σ^2, so that the statistic will have an approximate χ^2-distribution with N degrees of freedom, thus

$$\chi^2_{\text{obs}} = \sum_{j=1}^{N} \frac{(W_j - \mu)^2}{\sigma^2}. \tag{27}$$

Analyzed P-value has the form

$$P\text{-value} = \text{igamc}\left(\frac{N}{2}, \frac{\chi^2_{\text{obs}}}{2}\right). \tag{28}$$

If the obtained P-value is lower then accepted threshold (0.01) then the sequence is non-random as it has irregular occurrences of the possible patterns.

It is recommended to use patterns of length $m = 9$ or $m = 10$, the number o subblocks N should be choosen as $N \leq 100$ for validity of P-value and M should be choosen to satisfy $M > 0.01n$ and $N = [n/M]$

H. Overlapping template matching test

Based on [2, 20–22].

Overlapping template matching test is used to identifie non-random sequences which show too many or too few occurances of runs of ones of length m. It is possible to modify this test to detect irregular occurrences of any periodic pattern B.

The analyzed bit sequence, of length n, is divided into N blocks, each of length M, such that $n = MN$.

The number of runs of ones of length m, which possibly are overlapping in the j-th block is defined as $\tilde{W}_j = \tilde{W}_j(m, M)$. Analyzing the behaviour of random variable $\tilde{W}_j$ sequence, one can find that it converges in distribution the compound Poisson distribution of this random variable. If $(M - m + 1)\, 2^{-m} \to \lambda > 0$, then for real variable t

$$Ee^{t\tilde{W}_j} \to e^{\frac{\lambda e^t - 1}{2 - e^t}}. \tag{29}$$

The probabilites corresponding to above can be expressed with use of confluent hypergeometric function $\Phi = {}_1F_1$. Let U denotes some random variable with the compound Poisson asymtotic distribution, then for $u \leq 1$ and $\eta = \frac{\lambda}{2}$, the probability can be written as

$$P(U = u) = \frac{e^{-\eta}}{2^u} \sum_{l=1}^{u} \binom{u-1}{l-1} = \frac{\eta e^{-2\eta}}{2^u} \Phi(u + 1, 2, \eta). \tag{30}$$

From above one can easly calculate probabilites for small u, e.g.:

$$\begin{aligned} &P(U = 0) = e^{-\eta}, \quad P(U = 1) = \frac{\eta}{2} e^{-\eta}, \quad P(U = 2) = \frac{\eta}{8} e^{-\eta}(\eta + 2), \quad P(U = 3) = \frac{\eta}{8} e^{-\eta}\left(\frac{\eta^2}{6} + \eta + 1\right), \\ &P(U = 4) = \frac{\eta}{16} e^{-\eta}\left(\frac{\eta^3}{24} + \frac{\eta^2}{2} + \frac{3\eta}{2} + 1\right). \end{aligned} \tag{31}$$

Probabilites can also be calculated in the following manner [22].

- Consider a binary sequence of length $n - 1$ not matching an m-bit pattern B at all. In case of $n \leq m - 1$ then two sequences created by adding to the initial sequence zero or one at the end (both of length n) do not match the pattern B at all. If $n \geq m$ and the sequence of length $n - 1$ has an m-bit pattern in form of $011 \cdots 11$ at the tail of the sequence, then the sequence of length n created by adding one a the tail of the senquen matches the pattern B exactly once. The number of such sequences if $T_0(n - m - 1)$, which leads to recurrence formula for $T_0(n)$ as follows:

$$T_0(n) = \begin{cases} 1, & n = -1, \\ 1, & n = 0, \\ 2T_0(n-1), & 1 \leq n \leq m - 1, \\ 2T_0(n-1) - T_0(n - m - 1), & n \geq m. \end{cases} \tag{32}$$

- Consider a binary sequence of length n that matches pattern B exaclty once. Such sequences have bit pattern $011 \cdots 110$ of length $m + 2$. In case when pattern begins from the j-th bit of the sequence, there must by $T_0(j)$ patterns befor it and $T_0(n - m - 2 - j)$ patterns after it. Thus the recurrence formula for $T_1(n)$ is in form

$$T_1(n) = \begin{cases} 0, & n \leq m - 1, \\ 1, & n = m, \\ 2, & n = m + 1, \\ \sum_{j=-1}^{n-m-1} T_0(j)\, T_0(n - m - 2 - j), & n \geq m + 2. \end{cases} \tag{33}$$

- Some part of binary sequences of length n in which pattern B of length m occured exactly α times can be constructed by adding bit one at the first matching run of ones of binary sequence of length $n-1$ which has exactly $\alpha-1$ occurances of B, thus all the seqneces obtained in this manner will match the pattern B more then once at the first matchin run of ones. The number of such sequences is $T_{\alpha-1}(n-1)$.

  The rest of the binary sequences of length n in which pattern B of length m occured exactly α times match pattern B only once at the first matching run of ones. Consider a m-bit sequence which has $(m+2)$ bit pattern $011\cdots110$ with exactly zero occurrances of B befor it and exactly $\alpha-1$ occurrences of B after it. In case when pattern begins from the j-th bit of the sequence, the number of such sequences is $T_0(j)\,T_{\alpha-1}(n-m-2-j)$

  This leads to recurrence formula for $T_\alpha(n)$

$$T_\alpha(n)=\sum_{j=1}^{n-(m+\alpha-2)}T_0(j-2)\,T_0(n-(j+m+\alpha-1))+\sum_{k=1}^{\alpha=1}\sum_{j=1}^{n-2m-\alpha+2}T_0(j-2)\,T_k(n-(j+m+\alpha-k-1)), \tag{34}$$

  where $\alpha\geq 2$.

- Above result with simplified formula

$$T_\alpha(n)=T_{\alpha-1}(n-1)+\sum_{j=-1}^{n-2m-\alpha}T_0(j)\,T_{\alpha-1}(n-m-2-j), \tag{35}$$

  what allows to calculate the probabilities accordingly

$$\pi_i=\frac{T_i(n)}{2^n}, i=0,1,2,3,4,\quad \pi_5=1-\sum_{i=0}^{4}\pi_i. \tag{36}$$

The probabilites π_i, for $i=0,\ldots,5$, calculated with use of both metods are presented in table III.

	NIST values	[22] values
π_0	0.367879	0.364091
π_1	0.183940	0.185659
π_2	0.137955	0.139381
π_3	0.099634	0.100571
π_4	0.069935	0.0704323
π_5	0.140657	0.139865

TABLE III: Comparison of probabilites π_i which can be used for overlapping template matching test

The complement to the distribution function of analyzed random variable can be written as

$$L(u)=P(U>u)=e^{-\eta}\sum_{l=1}^{u}\frac{\eta^l}{l!}\sum_{k=1}^{u}\frac{1}{2^k}\binom{k-1}{l-1}. \tag{37}$$

To define the statistics $K+1$ classes should be choosen, i.e. $\{U=0\},\{U=1\},\ldots,\{U=K-1\},\{U\geq K\}$, and corresponding probabilites π_i should be calculated. Exemplary values: $K=2,\lambda=2,\eta=1$.

For m-bit pattern B one needs to calcuate the frequencies of occurrences ν_i, where ν_0 corresponds to none occurrences of pattern, ν_1 corresponds to 1 occurrences of pattern, etc. The frequencies are calculated by moving a frame of size m bit by bit in analyzed block and counting the occurrences of m length ones sequences. After analyzing of whole block the total number of occurrences indicates which class/frequency should be incremented.

Using calculated frequencies, and probabilites χ^2 statistics can be defined

$$\chi^2=\sum_{i=0}^{K}\frac{(\nu_i-N\pi_i)^2}{N\pi_i} \tag{38}$$

Analyzed P-value has the form

$$P\text{-value}=\text{igamc}\left(\frac{K}{2},\frac{\chi^2_{\text{obs}}}{2}\right). \tag{39}$$

If the obtained P-value is lower then accepted threshold (0.01) then the sequence is non-random as it has irregular occurrences of the possible patterns.

It is recommended

- to choose K, M, N that each sequence has length at least 10^6,
- m should have value 9 or 10,
- for other values of m one should take into account: $n \geq MN$, $N\,(\min \pi_i) > K$, $\lambda = \frac{(M-m+1)}{2^m} \approx 2$, $m \approx \log_2 M$, $K \approx 2\lambda$.

I. Maurers Universal Statistical Test

Based on [2, 23–28].

In 1992 Ueli Maurer (Princeton University) presented a test based on the statistic related with the per-bit entropy of the stream, described by author as the correct quality measure for a secret-key source for cryptography, allowing for measuring the value of key randomness defect in terms of running time of an enemy's optimal key-search stratego, i.e. effective key size of a cipher system.

The author states that this test is designed to detect any of a very general class of statistical defects that can be modeled by an ergodic stationary source with finit memory, this replaces a number of standard statistical tests.

The main idea is to detect whether or not the analyzed sequence can be significantly compressed without loss of information. If the sequence can be significantly compressed this signalize it is not truly random. It is based on the Ziv idea that a universal statistics test can be constructed on a universal source coding algorith. The Lempel-Ziv source-coding algorithm is considered by Maurer to be less suited for a statistical approach due to problem with difficulties with defining a test statistic which distribution could be analyzed.

The required length of an analyzed sequence for this test is quite large – $10 \cdot 2^L + 1000 \cdot 2^L, 6 \leq L \leq 16$. Such sequence is divided into

- two L-bit blocks, $6 \leq L \leq 16$,
- $Q \geq 10 \cdot 2^L$ initialization blocks, Q should be choosen to allow all possible L-bit patterns could occur within the initialization blocks,
- $K \approx 1000 \cdot 2^L$ test blocks, $K = \lceil \frac{n}{L} \rceil - Q$.

Too large L values are not recommended as the test initialization takes exponential time in L.

Test scheme can be prestented as follows

- initial sequence is being analyzed – walking through the sequence and analyzing L-bit blocks in test segments;
- during the walk, the test looks back on the entire sequence to find the nearest previous occurence of analyzed L-bit pattern;
- distance between this occuerence is recoreded;
- the algorithm calculates $\log_2$ of all recoreded distantces fo all L-bit patterns in test segment;
- results are averaged over all the expansion lengths by the number of test blocks.

With use of the initialization segment, a table for statistics definition is created

	...	binary value of ith pattern of length L	binary value of $(i+1)$th pattern of length L	...
	...	saved in $T_{j(i)}$	saved in $T_{j(i+1)}$	...
Initialization	...	last occurence index	last occurence index	...
Test block 1	...	last occurence index	last occurence index	...
Test block 2	...	last occurence index	last occurence index	...
Test block ...	...	last occurence index	last occurence index	...

TABLE IV: Exemplary table for $T_{j(i)}$ values, where $j\,(i)$ is the decimal value of the number i

$$f_n = \frac{1}{K} \sum_{i=Q+1}^{Q+K} \log_2 \left(i - T_j \left(i \right) \right), \tag{40}$$

where $T_j\left(i\right)$ is the table entry corresponding to the decimal representation of the contents of the i-th block of length L.

To performe this procedures in an efficient manner the algorithm is using a dynamic look-up table. The relation

$$\mathrm{E}\left(f_n\right) = 2^{-L} \sum_{i=1}^{\infty} \left(1 - 2^{-L}\right)^{i-1} \log_2 i, \tag{41}$$

where f_n is the test statistics, with expected value equal to expected value of a random variable $\log_2 G_L$ and G_L is a geometric random variable with paramenter $1 - 2^{-L}$.

It is possible to approximate the empirical formulas for the variance in several ways, for e.g.

$$\mathrm{Var}\left(f_n\right) = c\left(L, K\right) \mathrm{Var}\left(\log_2 G\right) / K, \tag{42}$$

where $c\left(L, K\right)$ contains the information about the occurences of the patterns in the analyzed sequence.

The initial sequence is at first partitioned into $r \leq 20$ substrings, and on each substring the value of the universal test statistics is calculated (for the same value of parameters K, L, Q). The P-value has the form

$$P\text{-value} = \mathrm{erfc}\left(\left| \frac{f_n - \mathrm{E}\left(L\right)}{\sqrt{\mathrm{Var}\left(f_n\right)}} \right| \right). \tag{43}$$

Or in approximation

$$P\text{-value} = \mathrm{erfc}\left(\left| \frac{f_n - \mathrm{E}\left(L\right)}{\sqrt{2) c \sqrt{\frac{\mathrm{Var}\, L}{K}}}} \right| \right), \tag{44}$$

where $c = 0.7 - \frac{0.8}{L} + \left(4 + \frac{32}{L}\right) \frac{K^{-\frac{3}{L}}}{15}$.

It is recommended to apply this test to a sequence of length at least n bits, where $n \geq \left(Q + K\right) L$.

If the obtained P-value is lower then accepted threshold (0.01) then the sequence is non-random.

J. Linear complexity test

Based on [2, 25, 26, 29].

In this test the linear complexity is used to verify randomness of analyzed sequence. The idea is based on so called Linear Feedback Shift Registers (LFSR), which are registers of length L consisting of L delay elements, each with single input and output. For example, if the initial state of LFSR is $\varepsilon_{L-1}, \ldots, \varepsilon_1, \varepsilon_0$ then output sequence is in form $\varepsilon_L, \varepsilon L_1, \ldots, \varepsilon_1$ where

$$j \geq L \quad \varepsilon_j = \left(c_1 \varepsilon_{j-1} + c_2 \varepsilon_{j-2} + \ldots + c_L \varepsilon_{j-L}\right) \mod 2, \tag{45}$$

with $c_1, \ldots, c_L$ are coefficients of the connection polynomial specific to a given LFSR. If some binary sequence is an output of LFSR for some initial state, it is said that it was generated by LFSR.

The linear complexity, $L\left(s^n\right)$, for some given sequence $s^n = \left(\varepsilon_1, \ldots, \varepsilon_n\right)$, is defined as the shortest LFSR which will generate s^n as its first n terms.

With use of Berlekamp-Massey algorithm it is possible to use the linear complexity for analyzing the randomness.

For a truly random binary sequence of length n, s^n, the mean and variance are defiend as follows

$$\begin{aligned} \mathrm{E}\left(L\left(s^n\right)\right) &= \frac{n}{2} + \frac{4 + B\left(n\right)}{18} - \frac{1}{2^n}\left(\frac{n}{3} + \frac{2}{9}\right), \\ \mathrm{Var}\left(L\left(s^n\right)\right) = \sigma_n^2 &= \frac{86}{81} - \frac{1}{2^n}\left(\frac{14 - B\left(n\right)}{27} n + \frac{82 - 2B\left(n\right)}{81}\right) - \frac{1}{2^{2^n}}\left(\frac{1}{9} n^2 + \frac{4}{27} n + \frac{4}{81}\right). \end{aligned} \tag{46}$$

Despite the suggestions in the Crypt-X package [25] the asymptotic distribution of $\frac{(L_n-\mu_n)}{\sigma_n}$ along the sequence of even or odd values of n defined on mixture of two geometric random variables does not exists, [2]. And thus the cases of n even and odd must be analyzed separately resulting in two different distributions in limits.

Due to this fack the following sequence of statistics is proposed

$$\begin{aligned} T_n &= (-1)^n (L_n - \xi_n) + \frac{2}{9} \\ \xi_n &= \frac{n}{2} + \frac{4 + r_n}{18}. \end{aligned} \tag{47}$$

The statistics from above take only integer values and coverge to the random variable T in distribution, which is skewed to the right.

$$\begin{aligned} P(T = 0) &= \frac{1}{2}, \quad k = 1, 2, \dots \\ P(T = k) &= \frac{1}{2^{2k}} \\ P(T = k) &= \frac{1}{2^{2|k|+1}}, \quad k = -1, -2, \dots \\ P(T \geq k > 0) &= \frac{1}{3 \cdot 2^{2k-2}} \\ P(T \leq k) &= \frac{1}{3 \cdot 2^{2|k|-1}} \end{aligned} \tag{48}$$

Thus allowing for evaluation of the P-value for the observerd T_{obs}.

$$P\text{-value} = \frac{1}{3 \cdot 2^{2\kappa-1}} + \frac{1}{3 \cdot 2^{2\kappa-2}} = \frac{1}{2^{2\kappa-1}} \tag{49}$$

As this distribution cannot attain the uniform character for P-values and has a discrete nature, the analyzed sequence must be of length $n = MN$, where N is the number of substrings of length M. For each of N substrings the statistic T_M is calculated, and $K + 1$ classes, depending on M, are selected. For each of analyzed substrings, the frequences, $\nu_0, \nu_1, \dots, \nu_K$ are evaluated upon the belonging of T_M to an apropriate class (one of $K + 1$).

Theoretical probabilites $\pi_0, \pi_1, \dots, \pi_K$ corresponding to $K + 1$ classes are determined with use of $P(T = k)$ for $k = 1, 2, \dots$ and $k = -1, -2, \dots$, in limitis for large enough M ($500 < M < 5000$).

The χ^2 statistic, with K degrees of freedom is constructed, in an approximation,

$$\chi^2 = \sum_{i=0}^{K} \frac{(\nu_i - N\pi_i)^2}{N\pi_i}. \tag{50}$$

The P-value is defined as follows

$$\frac{\int_{\chi^2_{\text{obs}}}^{\infty} e^{-\frac{u}{2}} u^{\frac{K}{2}-1} du}{\Gamma\left(\frac{K}{2}\right) 2^{\frac{K}{2}}} = \text{igamc}\left(\frac{K}{2}, \frac{\chi^2_{\text{obs}}}{2}\right) \tag{51}$$

The condition for above approximation is $N(\min \pi_i) \geq K$.

For recommended, sufficiently large values of M and N, one can propose six classes with apropriate probabilities:

- $\{T \leq -2.5\}$, the frequency ν_0, and the probability $\pi_0 = 0.010417$,
- $\{-2.5 < T \leq -1.5\}$, the frequency ν_1, and the probability $\pi_0 = 0.03125$,
- $\{-1.5 < T \leq -0.5\}$, the frequency ν_2, and the probability $\pi_0 = 0.125$,
- $\{-0.5 < T \leq 0.5\}$, the frequency ν_3, and the probability $\pi_0 = 0.5$,
- $\{0.5 < T \leq 1.5\}$, the frequency ν_4, and the probability $\pi_0 = 0.25$,
- $\{1.5 < T \leq 2.5\}$, the frequency ν_5, and the probability $\pi_0 = 0.0625$,
- $\{T > 2.5\}$, the frequency ν_6, and the probability $\pi_0 = 0.020833$.

It is recommended to choose $n \geq 10^6$, $500 \leq M \leq 5000$, and $N \geq 200$ for validity of the χ^2 approxiamtion.

If the obtained P-value is lower then accepted threshold (0.01) then the sequence is non-random

K. Serial test

Based on [2, 7, 26, 30, 31].

In this test the initial sequence is analyzed with use of variety of procedures baes on testing the uniformity of the distributions of patterns with given length. Basicly the test evaluates the frequencies of all possible overlapping patterns of length M in entire seqnece and verifies them with results for theoretical truly random sequence, which has uniformity – each pattern of given lenght has the same probability of occuerence as any other of the same lenght. In case of $m = 1$ this test becomes the frequency test described here as first.

Let the m-bit pattern is defined as $i_1, \ldots, i_m$, the initial sequence of length n is extended by its $m-1$ first bits giving the so called circularized sequence $\varepsilon_1, \ldots, \varepsilon_n, \varepsilon_1, \ldots, \varepsilon_{m-1}$, and $\nu_{i_1 \ldots i_m}$ is the frequency the occuerences of the pattern $i_1, \ldots, i_m$ in extended sequence.

A χ^2-type statistic ψ_m^2 is defined as follows

$$\begin{aligned}
\psi_m^2 &= \frac{2^m}{n} \sum_{i_1 \ldots i_m} \left(\nu_{i_1 \ldots i_m} - \frac{n}{2^m}\right)^2 = \frac{2^m}{n} \sum_{i_1 \ldots i_m} \nu_{i_1 \ldots i_m}^2 - n, \\
\psi_{m-1}^2 &= \frac{2^{m-1}}{n} \sum_{i_1 \ldots i_{m-1}} \left(\nu_{i_1 \ldots i_{m-1}} - \frac{n}{2^{m-1}}\right)^2 = \frac{2^{m-1}}{n} \sum_{i_1 \ldots i_{m-1}} \nu_{i_1 \ldots i_{m-1}}^2 - n, \\
\psi_{m-2}^2 &= \frac{2^{m-2}}{n} \sum_{i_1 \ldots i_{m-2}} \left(\nu_{i_1 \ldots i_{m-2}} - \frac{n}{2^{m-2}}\right)^2 = \frac{2^{m-2}}{n} \sum_{i_1 \ldots i_{m-2}} \nu_{i_1 \ldots i_{m-2}}^2 - n
\end{aligned} \tag{52}$$

but it does not have a χ^2 distribution, as the frequencies $\nu_{i_1 \ldots i_m}$ are not independent.

The generalized serial statistics can be written as

$$\begin{aligned}
\nabla \psi_m^2 &= \psi_m^2 - \psi_{m-1}^2 \\
\nabla^2 \psi_m^2 &= \psi_m^2 - 2\psi_{m-1}^2 + \psi_{m-2}^2,
\end{aligned} \tag{53}$$

where $\psi_0^2 = \psi_{-1}^2 = 0$.

Such statistic, $\nabla \psi_m^2$ has a χ^2 distribution with 2^{m-1} degrees of freedom, and $\nabla^2 \psi_m^2$ has a χ^2 distribution with 2^{m-2} degrees of freedom.

For recommended $m \leq \lfloor log_2(n) \rfloor - 2$ the P-values have the forms

$$\begin{aligned}
P\text{-value}_1 &= \text{igamc}\left(2^{m-2}, \frac{\nabla \psi_m^2}{2}\right) \\
P\text{-value}_2 &= \text{igamc}\left(2^{m-3}, \frac{\nabla^2 \psi_m^2}{2}\right)
\end{aligned} \tag{54}$$

If any of the obtained P-values is lower then accepted threshold (0.01) then the sequence is non-random.

L. Approximate entropy test

Based on [2, 32–34].

This test is also based on analyzing repeating occuenrces of given patterns in the initial sequence. All possible overlappings of patterns across the whole sequence are analyzed in scope of two blocks of consecutive lengts m and $m+1$ and compaired to results expected for truly random sequence.

Pattern of length m is defined as

$$Y_i(m) = (\varepsilon_i, \ldots, \varepsilon_{i+m-1}), \tag{55}$$

where $1 \leq i \leq n - m + 1$

The relative frequency of occurences of pattern $Y_i(m)$ in analyzed sequence has the form

$$C_i^m = \frac{1}{n+1-m} \{\text{number of such } j \text{ that } 1 \leq j < n-m, Y_j(m) = Y_i(m)\} = \pi_i. \tag{56}$$

And the entropy of the empirical distribution corresponding to all 2^m possible patterns of length m can be written as

$$-\Phi^{(m)} = -\frac{1}{n+1-m}\sum_{i+1}^{n+1-m}\log C_i^m, \tag{57}$$

thus

$$-\Phi^{(m)} = -\sum_{l=1}^{2^m}\pi_l \log \pi_l, \tag{58}$$

where π_l is the relative frequency of a pattern $l = i_1, \ldots, i_m$ in analyzed sequence.

Let $H(m)$ corresponds to the approximated entropy of an order $m \geq 1$, and be defined as

$$H(m) = \Phi^{(m)} - \Phi^{(m+1)}, \tag{59}$$

where $H(0) = -\Phi^{(m)}$. This approximated entropy is a measure for frequency describing the situation when blocks of length m which are close together remain close together when their lengths are extended by one bit. The extreme values of this entropy indicate deviation from truly random character – small values indictate strong regularity within the sequence and large values indicate fluctuations and irregularities.

In case when approximate entropy $H(m)$ has the largest possible value then corresponding sequence is called m-irregular or m-random [33]. Pincus and Kalman obtained quite interesting result while analyzing binary and decimal expansions of $e, \pi, \sqrt{2}, \sqrt{3}$ for $m = 1, 2, 3$ – $\sqrt{3}$ expansion has higher regularity than expansion of π.

This test is based on the fact that for a long random, and thus irregular, sequence, and for a fixed length m of a pattern, $H(m) \approx \log 2$. Rukhin showed the correlation between the limiting distribution $n[\log 2 - H(m)]$ and a χ^2 distribution for a variable with 2^m degrees of freedom.

This allows to define

$$\begin{aligned} \chi^2_{\text{obs}} &= n\left[\log 2 - H(m)\right], \\ P\text{-value} &= \text{igamc}\left(2^{m-1}, \frac{\chi^2_{\text{obs}}}{2}\right). \end{aligned} \tag{60}$$

It is possible to more precisely analyze this matter by modifying the definition of the approxiamte entropy.

For relative frequency $\nu_{i_1\ldots i_m} = \frac{\omega_{i_1\ldots i_m}}{n}$ of the pattern $i_1 \ldots i_m$ in extended sequence, $\varepsilon_1, \ldots, \varepsilon_n, \varepsilon_1, \ldots, \varepsilon_{m-1}$, one can write

$$\tilde{\Phi}^{(m)} = \sum_{i_1\ldots i_m} \nu_{i_1\ldots i_m}\log \nu_{i_1\ldots i_m}. \tag{61}$$

Thus $\omega_{i_1\ldots i_m} = \sum_k \omega_{i_1\ldots i_m k}$ resulting in $\sum_{i_1\ldots i_m}\omega_{i_1\ldots i_m} = n$ for any m.

Modified approximate entropy is defined as follows

$$\tilde{H}(m) = \tilde{\Phi}^{(m)} - \tilde{\Phi}^{(m+1)}. \tag{62}$$

As $\log s \geq \tilde{H}(m)$ for any m, due to Jensen's equality, it is possible that $\log s < \tilde{H}(m)$ which results in the largest possible value of $H(m) = \log s$ and its attained when $n = s^m$ and uniform distribution of all patterns of length m.

For large n approximate entropy and modified approximate entropy cannot differ much.

$$\begin{aligned} &\nu'_{i_1\ldots i_m} = \frac{\omega'_{i_1\ldots i_m}}{n-m+1}, \\ &\sum_{i_1\ldots i_m}\omega'_{i_1\ldots i_m} = n-m+1, \\ &\omega_{i_1\ldots i_m} - \omega'_{i_1\ldots i_m} \leq m-1, \\ &\Rightarrow \left|\nu_{i_1\ldots i_m} - \nu'_{i_1\ldots i_m}\right| \leq \frac{m-1}{n-m+1}. \end{aligned} \tag{63}$$

Which suggest that for given m, $\Phi^{(m)}$ and $\tilde{\Phi}^{(m)}$ must be close for large n. Which leads to closeness of Pincus approximate entropy and modified approximate entropy, and finally to coincidence of their asymptotic distributions.

If obtained P-value is lower then accepted threshold (0.01) then the sequence is considered as non-random.

It is recommended to choose m and n such that $m < \lfloor \log_2 n \rfloor - 5$.

M. Cumulative sums test – cusum

Based on [2, 12, 35].

In this test the initial binary sequence transformed to sequence of ones and minus ones (transformed from 0), is analyzed in scope of maximum absolute values of the partial sums. If the values are large then it indicates too many of ones or zeros (minus ones after transformation) in the early stages of the analyzed sequence, and on the other hand if the values are too small this indicates that mixing of ones and zeroes is too uniform.

It is possible to derive a similar test from the reversed time random walk with $S_k^{'} = X_n + \ldots + X_{n-k+1}$, with one difference – the results corresponding to the early stages now will correspond to the late stages of analyzed sequence.

The statistical concept of this test utilizes the limiting distribution of the maximum of the absoulte values of the partial sums

$$\max_{1\leq k\leq n} |S_k| \,, \tag{64}$$

which can be written as

$$\lim_{n\to\infty} P\left(\frac{\max\limits_{1\leq k\leq n}|S_k|}{\sqrt{n}} \leq z\right) = \frac{1}{\sqrt{2n}} \int_{k=-\infty}^{\infty} (-1)^k e^{-\frac{(u-2kz)^2}{2}} du = \frac{4}{\pi} \sum_{j=0}^{\infty} \frac{(-1)^j}{2j+1} e^{-\frac{(2j+1)^2\pi^2}{8z^2}} = H(z) \,, \tag{65}$$

where $z > 0$.

If the test statistics $z = \frac{\max\limits_{1\leq k\leq n}|S_k|_{\text{obs}}}{\sqrt{n}}$, attains large values then such sequence is rejected as non-random.

P-value for above statistics is defined as

$$P\text{-value} = 1 - H\left(\frac{\max\limits_{1\leq k\leq n}|S_k|_{\text{obs}}}{\sqrt{n}}\right) = 1 - G\left(\frac{\max\limits_{1\leq k\leq n}|S_k|_{\text{obs}}}{\sqrt{n}}\right), \tag{66}$$

where $G(z)$ is defined below.

As $H(z)$ converges quickly, it is convenient to use function $G(z)$ instead of $H(z)$ for moderate or large values of z and function $H(z)$ for small values. $G(z)$ is equal to $H(z)$ for all z.

$$\begin{aligned} G(z) &= \frac{1}{\sqrt{2\pi}} \int_{-z}^{z} \sum_{k=-\infty}^{\infty} (-1)^k e^{-\frac{(u-2kz)^2}{2}} du = \sum_{k=-\infty}^{\infty} (-1)^k \left(\Phi\left((2k+1)z\right) - \Phi\left((2k-1)z\right)\right) \\ &= \Phi(z) - \Phi(-z) + 2\sum_{k=1}^{\infty} (-1)^k \left(\Phi\left((2k+1)z\right) - \Phi\left((2k-1)z\right)\right) \\ &= \Phi(z) - \Phi(-z) - 2\sum_{k=1}^{\infty} \left(2\Phi\left((4k-1)z\right) - \Phi\left((4k+1)z\right) - \Phi\left((4k-3)z\right)\right) \\ &\approx \Phi(z) - \Phi(-z) - 2\left(2\Phi(3z) - \Phi(5z) - \Phi(z)\right) \approx 1 - \frac{4}{\sqrt{2\pi}z} e^{-\frac{z^2}{2}}, \quad z\to\infty, \end{aligned} \tag{67}$$

where $\Phi(x)$ is a standard normal distribution.

Using the following theorem (Theorem 2.6 from [12])

Theorem For any integers $a \leq 0 \leq b$ and $a \leq u \leq \nu \leq b$ one has

$$\begin{aligned} P\left(a < -M_n^- \leq M_n^+ < b, u < S_n < \nu\right) = &\sum_{k=-\infty}^{\infty} P\left(u + 2k(b-a) < S_n < \nu + 2k(b-a)\right) \\ &- \sum_{k=-\infty}^{\infty} P\left(2b - \nu + 2k(b-a) < S_n < 2b - u + 2k(b-a)\right) \end{aligned} \tag{68}$$

$$\begin{aligned} P\left(a < -M_n^- \leq M_n^+ < b\right) = &\sum_{k=-\infty}^{\infty} P\left(a + 2k(b-a) < S_n < b + 2k(b-a)\right) \\ &- \sum_{k=-\infty}^{\infty} P\left(b + 2k(b-a) < S_n < 2b - a + 2k(b-a)\right) \end{aligned} \tag{69}$$

$$P\left(M_n < b\right) = \sum_{k=-\infty}^{\infty} P\left(\left(4k-1\right)b < S_n < \left(4k+1\right)b\right) - \sum_{k=-\infty}^{\infty} P\left(\left(4k+1\right)b < S_n < \left(4k+3\right)b\right) \tag{70}$$

one can obtain

$$P\left(\max_{1\le k\le n}|S_n| \ge z\right) = 1 - \sum_{k=-\infty}^{\infty} P\left(\left(4k-1\right)z < S_n < \left(4k+1\right)z\right) + \sum_{k=-\infty}^{\infty} P\left(\left(4k+1\right)z < S_n < \left(4k+3\right)z\right). \tag{71}$$

Above is used to calculate proper P-value for this test with $z = \frac{\max\limits_{1\le k\le n}|S_{k\,\mathrm{obs}}|}{\sqrt{n}}$.

If obtained P-value is lower then accepted threshold (0.01) then the sequence is considered as non-random.

It is recommended that $n \le 100$.

N. Random excursions test

Based on [2, 12, 35, 36].

This test analyzies the distribution of the number of visits to a given state as a result of an excurison of a simple random walk – the initial sequence is transformed to contain only ones and minus ones (instead of zeros) and successive sums of bits are considered.

Let $S_k = X_1 + \ldots + X_k$ dentoe a simple random walk, where X_i are independent variables taking values ± 1 with probabilites p and $1 - p = q$ accordingly.

For $S_0 = 0$ let $\rho_1 < \rho_2 < \ldots$ denote times when origin of the walk is reached

$$\begin{aligned} \rho_1 &= \min\left\{k, k > 0, S_k = 0\right\} \\ \rho_2 &= \min\left\{k, k > \rho_1, S_k = 0\right\} \\ &\ldots. \end{aligned} \tag{72}$$

With this random walk a sequence of excursions to and from zero can be associated in form

$$\left(S_0, \ldots, S_{\rho_1}\right), \left(S_{\rho_1}, \ldots, S_{\rho_2}\right), \ldots, \tag{73}$$

or

$$\left(i, \ldots, l\right) : S_{i-1} = S_{i+1} = 0, \quad S_k \ne 0, \quad i \le k \le l. \tag{74}$$

If J dentoes the total number of such excursions in the analyzed sequence – thus be a random variable – then the limiting distribution for it will be in form

$$\lim_{n\to\infty} P\left(\frac{J}{\sqrt{n}} < z\right) = \sqrt{\frac{2}{\pi}} \int_0^z e^{-\frac{u^2}{2}} du, \quad z > 0, \tag{75}$$

and corresponding P-value can be written as

$$P\left(J < J_{\mathrm{obs}}\right) \approx \sqrt{\frac{2}{\pi}} \sum_0^J e^{-\frac{u^2}{2}} du = P\left(\frac{1}{2}, \frac{J_{\mathrm{obs}}^2}{2n}\right). \tag{76}$$

For small values of J, $J < \max\left\{0.005\sqrt{n}, 500\right\}$, the analyzed sequence is considered to be non-random. For $J \ge \max\left\{0.005\sqrt{n}, 500\right\}$ the number of visists during the random walk in given state is calculated.

Let $\xi\left(x\right)$ dentoes the number of visits to the state $x \ne 0$, during a single first excursion. Its distribution can be derived as follows.

It can be derived that for $\xi\left(x\right) > 0$, $\xi\left(x\right)$ has a geometric distribution with parameter

$$\pi = \begin{cases} \frac{|p-q|}{1-\left(\frac{q}{p}\right)^x}, & p > q, x > 0 \text{ or } p < q, x < 0, \\ \frac{|p-q|}{1-\left(\frac{p}{q}\right)^x}, & p > q, x < 0 \text{ or } p < q, x > 0, \end{cases} \tag{77}$$

where $P\left(\xi\left(x\right)=0\right)=1-\frac{|p-q|}{\left|1-\left(\frac{p}{q}\right)^{x}\right|}$

Thus $\xi\left(x\right)$ has a zero-modified geometric distribution with probabilities,

$$\begin{aligned} &P\left(\xi\left(x\right)=0\right)=p_0,\\ &P\left(\xi\left(x\right)=k\right)=\left(1-p_0\right)\pi\left(1-\pi\right)^{k-1},\quad k=1,2,\ldots, \end{aligned} \tag{78}$$

resulting in

$$\begin{aligned} E\left(\xi\left(x\right)\right)&=\frac{1-p_0}{\pi},\\ \mathrm{Var}\left(\xi\left(x\right)\right)&=\frac{\left(1-p_0\right)\left(1-\pi+p_0\right)}{\pi^2}, \end{aligned} \tag{79}$$

and for $p=\frac{1}{2}$

$$\begin{aligned} &E\left(\xi\left(x\right)\right)=1,\\ &\mathrm{Var}\left(\xi\left(x\right)\right)=4\left|x\right|-2. \end{aligned} \tag{80}$$

For $a=0,1,2,\ldots$ one can write

$$\begin{aligned} P\left(\xi\left(x\right)>a\right)&=\left(1-p_0\right)\left(1-\pi\right)^{a}=\frac{P\left(\xi\left(x\right)=a+1\right)}{\pi},\\ p=\frac{1}{2}\rightarrow\quad P\left(\xi\left(x\right)>a\right)&=\frac{1}{2\left|x\right|}\left(1-\frac{1}{2\left|x\right|}\right)^{a}=2\left|x\right|P\left(\xi\left(x\right)=a+1\right). \end{aligned} \tag{81}$$

These results allow to test randomness of a bit sequence in following manner

- For given collection of states, x-values (for example $-4\leq x\leq 4, x\neq 0$ or $-7\leq x\leq 7, x\neq 0$), the observed frequencies of k visits of the state x during J excursions, $\nu_k\left(x\right)$, are evaluated.

$$\nu_k\left(x\right)=\sum_{j=1}^{J}\nu_k^j\left(x\right), \tag{82}$$

  where $\nu_k^j\left(x\right)=1$ if the state x is visited during the j-th excursion, $j=1,\ldots,J$, exactly k times, and $\nu_k^j\left(\right)=0$) otherwise.

- The values $\xi\left(x\right)$ are divided into classes with different k, for example $k=0,1,\ldots,4$ and $k\geq 5$.

- For above classes the theoretical probabilites are calculated as follows

$$\begin{aligned} \pi_0&=P\left(\xi\left(x\right)=0\right)=1-\frac{1}{2\left|x\right|},\\ \pi_k&=P\left(\xi\left(x\right)=k\right)=\frac{1}{4x^2}\left(1-\frac{1}{2\left|x\right|}\right)^{k-1},k=1,\ldots,4,\\ \pi_5&=P\left(\xi\left(x\right)\geq 5\right)=\frac{1}{2\left|x\right|}\left(1-\frac{1}{2\left|x\right|}\right)^{4}. \end{aligned} \tag{83}$$

For any x,

$$\chi^2\left(x\right)=\sum_{k=0}^{5}\frac{\left(\nu_k\left(x\right)-J\pi_k\left(x\right)\right)^2}{J\pi_k\left(x\right)}, \tag{84}$$

should has an approximated χ^2 distribution with 5 degrees of freedom, where $J\min\pi_k\left(x\right)\geq 5$, i.e. $J\geq 500$ – if not larger set of classes should be assumed.

P-values corresponding to above can be written as

$$1-P\left(\frac{5}{2},\frac{\chi^2_{\mathrm{obs}}\left(x\right)}{2}\right). \tag{85}$$

The whole procedure can be presetned in the following steps

1. Analyzed sequence is in form $\varepsilon_0\varepsilon_1\ldots\varepsilon_n$.

2. This sequence is transformed into X, $X_i = 2\varepsilon_i - 1$.

3. The set $S = \{S_i\}$ is calculated, where S_i is a partial sum, $S_i = \sum_{k=1}^{i} X_k$.

4. Creation of set $S^{'}$ by prepending set S with $S_0 = 0$ and appending set S with $S_n = 0$.

5. This generates a random walk.

6. J is the total number of zeros in $S^{'}$ without the starting zero – J in fact is the number of cycles in $S^{'}$.

7. For each cycle and for each x value, without zero, for example $-p \leq x \leq -1$ and $1 \leq x \leq p$, where p is test parameter, frequency of occurence of each x is calculated within each cycle (for example with use of table V).

	cycles		
state x	cycle_1	$\ldots$	cycle_J
$x = -p$			
$\ldots$			
$x = -1$			
$x = 1$			
$\ldots$			
$x = p$			

TABLE V: Occurences of certain state x within cycles.

8. For each state x the appropriate frequencies are stored in $\nu_k(x)$ – it is equal to the number of cycles in which givne state x occuers exactly k times, frequencies higher then k are stored in $\nu_k(x)$. $\sum_{l=0}^{k} \nu_l(x) = J$. The result can be presented in table VI.

	number of cycles		
state x	1	$\ldots$	J
$x = -p$			
$\ldots$			
$x = -1$			
$x = 1$			
$\ldots$			
$x = p$			

TABLE VI: Frequencies of certain state x within cycles.

9. For each state x the statistics is computed from the relation

$$\chi^2(x) = \sum_{l=0}^{k} \frac{(\nu_l(x) - J\pi_l(x))^2}{J\pi_l(x)}, \tag{86}$$

where the probabilites π_l are computed as theoretical what was presented above.

10. Finally the P-value is calculated for each state as

$$P\text{-value} = \text{igamc}\left(\frac{k}{2}, \frac{\chi^2_{\text{obs}}}{2}\right). \tag{87}$$

If obtained P-value is lower then accepted threshold (0.01) then the sequence is considered as non-random.

It is recommended that any analyzed sequence be of length $n \geq 10^6$ and $J \min \pi_k(x) \geq k$.

O. Random excursions variant test

Based on [2, 12, 35, 36].

Using the random excursion test definition one can derive its variant.

With use of $\xi(x)$ which was the number of visits of state x in the single first excursion let $\xi_J(x)$ be the total number of visits of state x during J excursions. As S_k renews itself at every occuerence of 0 in the excursion, $\xi_J(x)$ will be a sum of independent identically distributed variables with the same distribution as $\xi(x) = \xi_1(x)$, giving the relation for limiting distribution as follows $\xi_J(x)$

$$\lim_{J\to\infty} P\left(\frac{\xi_J(x) - J}{\sqrt{J(4|x| - 2)}} < z\right) = \Phi(z), \tag{88}$$

which in this case is a normal distribution.

Thus the corresponding P-value has the form

$$P\text{-value} = \operatorname{erfc}\left(\frac{|\xi_{J_{\text{obs}}}(x) - J|}{\sqrt{2J(4|x| - 2)}}\right). \tag{89}$$

III. DIEHARDER TESTS OVERVIEW

[1] J. von Neumann, *Mathematical Foundations of Quantum Mechanics* (Princeton Univ. Press, Princeton, 1955).
[2] A. Rukhin, J. Soto, J. Nechvatal, M. Smid, E. Barker, S. Leigh, M. Levenson, M. Vangel, D. Banks, A. Heckert, J. Dray, and S. Vo, NIST Special Publication 800-22 Revision 1a (2010), revisor: Lawrence E. Bassham III.
[3] K. L. Chung, *Elementary Probability Theory with Stochastic Processes* (Springer Verlag, New York, 1979).
[4] J. Pitman, *Probability* (Springer Verlag, New York, 1993).
[5] J. A. Rice, *Mathematical Statistics and Data Analysis (Second ed.)* (Duxbury Press, Belmont, 1995).
[6] N. Maclaren, *Cryptographic Pseudo-random Numbers in Simulation. Cambridge Security Workshop on Fast Software Encryption* (R. Anderson, Cambridge, 1993).
[7] D. E. Knuth, *The Art of Computer Programming. Vol 2: Seminumerical Algorithms. 3rd ed.* (Addison-Wesley, Reading, 1998).
[8] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions: NBS Applied Mathematics Series 55* (U.S. Government Printing Office, Washington, 1967).
[9] J. D. Gibbons, *Nonparametric Statistical Inference, 2nd ed.* (Marcel Dekker, New York, 1985).
[10] A. P. Godbole and S. G. Papastavridis, *Runs and patterns in probability: Selected papers* (Kluwer Academic, Dordrecht, 1994).
[11] F. N. David and D. E. Barton, *Combinatorial Chance* (Hafner Publishing Co., New York, 1962).
[12] P. Revesz, *Random Walk in Random and Non-Random Environments* (World Scientific, Singapore, 1990).
[13] G. Marsaglia, *DIEHARD: a battery of tests of randomness*.
[14] I. N. Kovalenko, Theory of Probability and its Applications **17**, 342 (1972).
[15] G. Marsaglia and L. H. Tsay, Linear Algebra and its Applications **67**, 147 (1985).
[16] R. N. Bracewell, *The Fourier Transform and Its Applications* (McGraw-Hill, New York, 1986).
[17] W. Killman, J. Schth, W. Thumser, and I. Uludag, T-Systems, Systems Integration (2004).
[18] S. Kim, K. Umeno, and A. Hasegawa, Cryptology ePrint Archive, Report 2004/018 (2004).
[19] A. D. Barbour, L. Holst, and S. Janson, *Poisson Approximation* (Clarendon Press, Oxford, 1992).
[20] O. Chrysaphinou and S. Papastavridis, Probability Theory and Related Fields **79**, 129 (1988).
[21] N. J. Johnson, S. Kotz, and A. Kemp, *Discrete Distributions. 2nd ed.* (John Wiley, New York, 1996).
[22] K. Hamano and T. Kaneko, IEICE Transactions of Electronics, Communications and Computer Sciences **E90-A**, 1788 (2007).
[23] U. M. Maurer, Journal of Cryptology **5**, 89 (1992).
[24] J.-S. Coron and D. Naccache, *An Accurate Evaluation of Maurer's Universal Test. Proceedings of SAC '98 (Lecture Notes in Computer Science)* (Springer Verlag, Berlin, 1998).
[25] H. Gustafson, E. Dawson, L. Nielsen, and W. Caelli, Computers & Security **13**, 687 (1994).
[26] A. J. Menezes, P. C. van Oorschot, and S. A. Vanstone, *Handbook of Applied Cryptography* (CRC Press, Boca Raton, 1997).
[27] J. Ziv, *Compression, tests for randomness and estimating the statistical model of an individual sequence. Sequences (ed. R.M. Capocelli)* (Springer Verlag, Berlin, 1990).

[28] J. Ziv and A. Lempel, IEEE Transactions on Information Theory **23**, 337 (1977).
[29] R. A. Rueppel, *Analysis and Design of Stream Ciphers* (Springer Verlag, New York, 1986).
[30] J. Ziv and A. Lempel, Proc. Cambridge Philos. Soc. **47**, 276 (1953).
[31] J. Ziv and A. Lempel, Electronics Letters **23**, 365 (1987).
[32] S. Pincus and B. H. Singer, Proc. Natl. Acad. Sci. USA **93**, 2083 (1996).
[33] S. Pincus and R. E. Kalman, Proc. Natl. Acad. Sci. USA **94**, 3513 (1997).
[34] A. Rukhin, Journal of Applied Probability **37**, 88 (2000).
[35] F. Spitzer, *Principles of Random Walk* (Van Nostrand, Princeton, 1964).
[36] M. Baron and A. L. Rukhin, Communications in Statistics: Stochastic Models **15**, 593 (1999).