A natural approach to the experimental study of dependence between statistical tests

A. M. Zubkov, A. A. Serov

Academy of Cryptography of the Russian Federation
Steklov Mathematical Institute of Russian Academy of Sciences

Moscow region 2020

Generators of random and pseudo-random sequences (RNG and PRNG) are of great value for cryptography.

According to the Shannon theory of secret communications the ideal ciphers should generate ciphertexts which are indistinguishable from the equiprobable Bernoulli sequence. Independence and uniformity of distribution of cryptographic keys maximize security with respect to brute force attack.

The testing of output binary sequences generated by (Pseudo) Random Number Generators aims to decide whether these sequences may be considered as realizations of equiprobable Bernoulli sequence and therefore may be used in cryptosystems.

There are several popular packages of statistical tests which are distributed with open source codes (e. g. TESTU01, DIEHARD, NIST, SPRNG), or with closed source codes (e. g. Crypt-X).

Each package consists of several tests based on more or less standard goodness-of-fit tests of mathematical statistics. Clearly, no finite collection of tests may be sufficient to prove exact correspondence of RNG or PRNG to the ideal Bernoulli source. Moreover, PRNG generates deterministic nonrandom sequences, and physical RNG are real objects which cannot be functions defined on measurable space of elementary events.

Exact correspondence may be superfluous for cryptography. The problem of choosing tests which are really necessary for cryptographic applications is important, but is very hard, especially in view of development of cryptanalytic methods.

NIST Statistical Tests

Test No.	Name of Test and character of statistics used
1	**Frequency** (of ones)
2	**Block Frequency** (of ones in 128-bit blocks)
3	**Runs** (number of sign changes)
4	**Longest Run** (1-runs in 10^4-bit blocks)
5	**Binary Matrix Rank** (ranks of binary $32x32$-matrices)
6	**Discrete Fourier Transform** (number of large coefficients)
7	**Overlapping Template Matching** (numbers of 9-bit 1-series in 1032-bit blocks)
8	**Universal (Maurer's "Universal statistical test")** (sum of logarithms of distances between equal 7-bit blocks)
9	**Linear Complexity** (for 500-bit blocks)
10	**Serial** (intersecting 16-, 15- and 14-bit blocks)
11	**Approximate Entropy** (intersecting 10- and 11-bit blocks)
12	**Cumulative Sums** (maximal deviation of random walk from 0)

Table: List of NIST Statistical Tests considered

Comments on the NIST Statistical package

We have excluded 3 tests from 15 tests in the NIST package: Non-overlapping Template Matching Test, Random Excursions and Random Excursions Variant tests.

The Non-overlapping Template Matching Test generates 148 p-values and corresponding decisions.
"It is highly likely" that among such large family of tests there exist dependent ones.

We have performed additional experiments showing that in the collection of 162=148+14 tests of NIST package there exist obvious dependencies. Corresponding results will be discussed at the end of this report.

Comments on the NIST Statistical package

NIST Statistical Suite recommends to discontinue the Random Excursions and Random Excursions Variant tests if the trajectory of ± 1 walk have smaller than 500 returns to 0. But for simple symmetrical random walk r-th return to 0 becomes at the even moment n with the probability

$$\varphi_{r,n} = \frac{r}{(n-r)2^{n-r}} C_{n-r}^{n/2} = r\sqrt{\frac{2}{\pi n^3}} e^{-\frac{r^2}{2n}} \left(1 + O\left(\frac{r}{n} + \frac{r^3}{n^2}\right)\right),$$

if $n \to \infty$, $r = \mathrm{const}$. The probability that there are smaller than 500 returns to 0 during $n = 10^6$ steps is larger than 0.3, and for $n = 10^7$ steps it is larger than 0.1. So, tests based on Random Excursions do not generate any p-value for large part of tested segments, and it is unreasonable to use such tests if the probability of error is chosen to be, for example, 0.01.

Comments on the NIST Statistical package

In the NIST Statistical Suite description the authors had reported that investigation of empirical distributions of p-values of statistics of the collection of tests was performed by means of:
the Kolmogorov–Smirnov test,
factor analysis of 161 statistics for different tests,
principal component analysis,
correlation matrix of transformed p-values.

The results were formulated as follows:
1) there is no large redundancy among the tests,
2) the degree of duplication among the tests seems to be very small.

Along with this summary it was noted that there exist dependencies between some tests.

Comments on the NIST Statistical package

The dependencies between tests of the NIST package were studied by Iwasaki (arxiv 1710.01441). He had applied the same 162 tests to m segments consisting of n bits each. For each segment the arithmetic mean of p-values was computed and the histogram of arithmetic means was compared with the normal density approximating the sum of 162 independent random variables with uniform distribution on the interval (0,1). It appears that the histogram is wider than normal density; this is interpreted as the existence of positive correlations between p-values.

Correlations between statistics or p-values of NIST tests were studied also by different authors, see, e.g., Burciu, Simeon (IEEE Contr. Comput. Sci., 2019, v.5, iss.15, p.1-6) and references therein.

We suggest to measure the dependencies between tests in terms of decisions made by tests, and apply this approach to tests of the NIST package.

This reduces testing the independence between tests to testing the independence of components of multidimensional binary vector. Further, since to estimate probabilities of values of the vector the sample size should be significantly larger than the number of its values, we use simplified natural statistics.

For each of $m = 10000$ tested segments we compute the number of tests rejecting it and compare frequencies of these m numbers with probabilities of binomial distribution $\mathrm{Bin}(m, q)$, where q is the rejection probability for the segment of equiprobable Bernoulli sequence. This binomial distribution corresponds to the case of independent decisions of tests.

Description of our experiments

We have applied 12 tests from Table 1 to pseudorandom sequences of two types:
a) sequences generated by the block cipher AES,
b) sequences obtained from pairs of binary linear recurrent sequences.

Properties of type b) sequences were reported at CTCrypt'18; it was shown (on small sample examples) that, as a rule, such sequences successfully pass all tests of the NIST package. Here we use samples of 10000 segments with length 2^{20} bit.

Description of our experiments

Sequences of type a) were obtained by means of AES block cipher in the CFB (Cipher Feedback Block) mode with the zero starting plaintext block, IV=0xffffffffffffffffffffffffffffff80, zero key for the sequence denoted by AES_0, key 0x2b7e151628aed2a6abf7158809cf4f3c for the sequence denoted by AES_1.

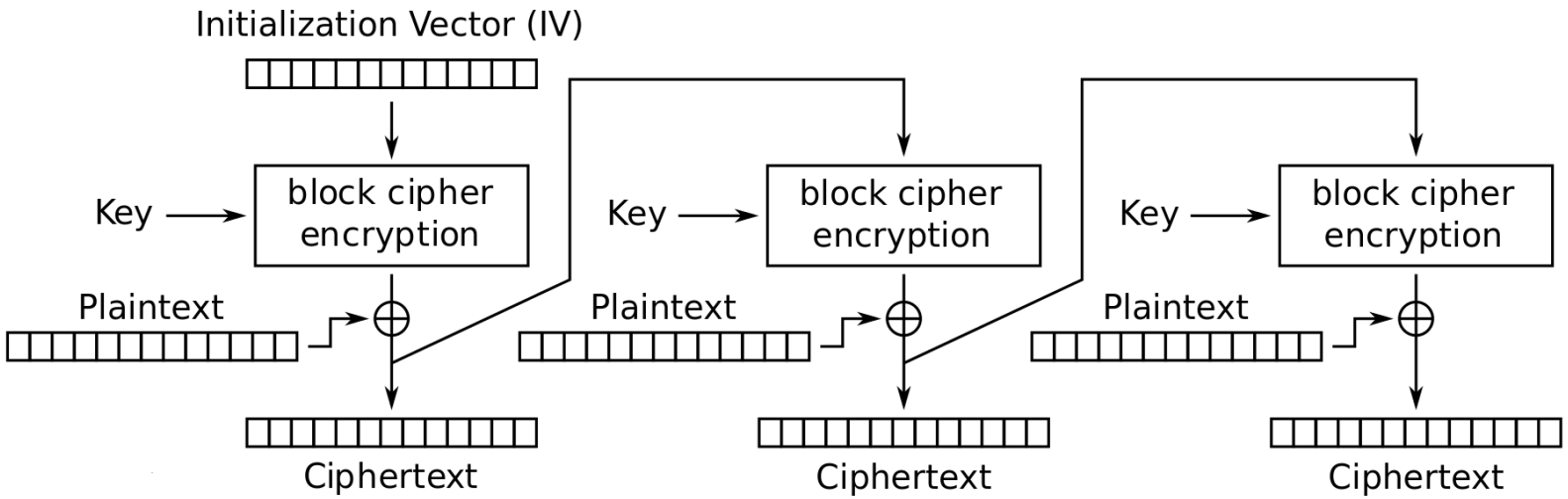

Figure:

Each sequence of type b) consists of concatenated segments of two linear recurrent sequences $\{x(i)\}_{i=0}^{2^{n_1}-1}$ and $\{z(i)\}_{i=0}^{2^{n_2}-1}$ over $\mathrm{GF}(2)$:
$\{x(1),\ldots,x(L_1^*),z(1),\ldots,z(L_2^*),$
$\quad x(L_1^*+1),\ldots,x(L_1^*+L_3^*),z(L_2^*+1),\ldots,z(L_2^*+L_4^*),\ldots\},$
where $x(1),\ldots,x(L_1^*)$ is the initial segment of sequence $\{x(i)\}$ with length $L_1^*=64$,
$z(1),\ldots,z(L_2^*)$ is the initial segment of sequence $\{z(i)\}$ with length

$$L_2^*=64+(x(L_1^*-5)x(L_1^*-4)x(L_1^*-3)x(L_1^*-2)x(L_1^*-1)x(L_1^*))_2,$$

$(x_{L_1^*+1},x_{L_1^*+2},\ldots,x_{L_1^*+L_3^*})$ is the segment of sequence $\{x_i\}$ with length

$$L_3^*=64+(z(L_2^*-5)z(L_2^*-4)z(L_2^*-3)z(L_2^*-2)z(L_2^*-1)z(L_2^*))_2,$$

and so on.

Tested sequences were constructed from linear recurrent sequences over GF(2) with maximal periods and characteristic polynomials
$f(x) = x^{43} + x^{27} + x^{22} + x^5 + 1$,
$g(x) = x^{63} + x + 1$,
$h(x) = x^{33} + x^{26} + x^{22} + x^{21} + x^3 + x + 1$,
$m(x) = x^{33} + x^{13} + 1$,
$q(x) = x^{63} + x^{60} + x^{22} + x^{18} + x^8 + x^5 + 1$,
$u(x) = x^{64} + x^{61} + x^{56} + x^{31} + x^{28} + x^{23} + 1$
by concatenating segments of following pairs:
MixL_1: $h(x)$, $m(y)$ (equal orders, 3-term, distant powers),
MixL_2: $f(x)$, $g(y)$ (different orders, 3-term, adjacent powers),
MixL_3: $f(x)$, $q(y)$ (different orders),
MixL_4: $u(x)$, $q(y)$ (different orders).

Description of our experiments

The application of 12 tests to each segment results in 14 different p-values (each of Serial and Cumulative Sums results in two p-values).

The critical level of p-value was chosen as 0.01: if the p-value does not exceed 0.01, then test rejects the hypothesis H_0 on the randomness and equiprobability of elements of tested segment.

For each segment we compute the number of p-values not exceeding 0.01, and for each tested sequence compute frequencies $\eta_0, \eta_1, \ldots, \eta_{14}$ of these numbers in sample of 10000 size.

Under the hypothesis H_{ind} on the independence of tests for each segment of "ideal" Bernoulli sequence the number of tests giving p-value not exceeding 0.01 should have the binomial distribution with the parameters $(14,\, 0.01)$. Since for each sequence we have considered $s = 10000$ segments, then for independent tests

$$\mathbf{E}\eta_k = 10000 C_{14}^k (0.01)^k (0.99)^{14-k}, \quad k = 0, 1, \ldots, 14.$$

To test the hypothesis H_{ind} we use the Pearson statistics for $(\eta_0, \eta_1, \eta_2 + \ldots + \eta_{14})$, joining all cases with more than one p-value not exceeding 0.01 in one class:

$$Pear = \sum_{k=0}^{1} \frac{(\eta_k - \mathbf{E}\eta_k)^2}{\mathbf{E}\eta_k} + \frac{\left(\sum_{k=2}^{14} \eta_k - \sum_{k=2}^{14} \mathbf{E}\eta_k\right)^2}{\sum_{k=2}^{14} \mathbf{E}\eta_k}.$$

Under hypothesis H_{ind} the distribution of this statistics should be close to chi-square distribution with 2 degrees of freedom having mean 2 and variance 4.

Table 2 contains frequencies $\eta_0, \eta_1, \ldots, \eta_9$ for samples of 10000 segments of sequences AES_0, AES_1, MixL_1, MixL_2, MixL_3, MixL_4.

Last column contains values $\mathbf{E}\eta_k$ for the binomial distribution with the parameters $(14,\, 0.01)$ (for the case of independent tests). Last row contains values of Pearson statistics $Pear$.

Table of frequencies of numbers of simultaneously rejecting tests of 14 ones

k	AES_0	AES_1	MixL_1	MixL_2	MixL_3	MixL_4	$\mathbf{E}\eta_k$
0	8768	8826	8729	5602	8816	8800	8687.46
1	1012	992	1103	1906	983	1023	1228.53
2	125	100	100	978	128	108	80.66
3	83	71	58	734	63	59	3.26
4	12	9	9	496	9	8	0.09
5	0	1	1	158	1	2	0.018
6	0	1	0	59	0	0	$2.7 \cdot 10^{-5}$
7	0	0	0	29	0	0	$3.2 \cdot 10^{-7}$
8	0	0	0	21	0	0	$2.8 \cdot 10^{-9}$
9	0	0	0	17	0	0	$1.9 \cdot 10^{-11}$
Total	10000	10000	10000	10000	10000	10000	10000
Pear	259	162	97	70487	214	139	

Table: Numbers of segments resulting in k (out of 14) p-values smaller than 0.01 and values of Pearson statistics for truncated binomial distribution

The table shows that the hypothesis on the independence of 14 tests from NIST Suite should be rejected.

Note that the sequence MixL_2 often is rejected by many tests. This may be a consequence of bad mixing properties of linear recurrent sequence with the 3-term characteristic polynomial $g(x) = x^{63} + x + 1$. (Nevertheless in more than half segments of MixL_2 all tests were passed successfully).

Further, p-values corresponding the Frequency, the Cumulative Sums and the Serial tests are simultaneously smaller than 0.01 with probabilities larger than for independent tests.

This effect may be justified theoretically. If $s_0 = 0, s_1, \ldots, s_n$ is the walk on $\mathbb{Z}$ with steps ± 1, then the number of walks with $s_n = 2w - n$ and $z_n = \max_{0 \leqslant k \leqslant n} |s_k| \leqslant u$ is equal to

$$L(w, n-w, u) = \sum\nolimits_{k \in \mathbb{Z}} \left(C_n^{w-2k(u+1)} - C_n^{w-(2k-1)(u+1)} \right).$$

Since s_n and z_n correspond to statistics of Frequency and Cumulative Sums tests, this equation permits to compute the correlations and joint distributions of these statistics: the correlation of $|s_n|$ and z_n is approximately 0.85 and

$$\mathbf{P}\{|s_n| > a, z_n > b\} \approx \min\{\mathbf{P}\{|s_n| > a\}, \mathbf{P}\{z_n > b\}\},$$

if $\mathbf{P}\{|s_n| > a\}$ and $\mathbf{P}\{z_n > b\}$ are smaller than 0.01.

Table of frequencies of numbers of simultaneously rejecting tests of 11 ones

k	AES_0	AES_1	MixL_1	MixL_2	MixL_3	MixL_4	$\mathbf{E}\eta_k$
0	8878	8920	8835	5650	8919	8905	8953.38
1	1043	1016	1097	2189	1012	1025	994.82
2	78	61	64	1212	66	66	50.24
3	1	2	4	638	3	4	1.52
4	0	1	0	202	0	0	0.03
5	0	0	0	67	0	0	$4 \cdot 10^{-4}$
6	0	0	0	25	0	0	$4.4 \cdot 10^{-6}$
7	0	0	0	13	0	0	$3.2 \cdot 10^{-8}$
8	0	0	0	4	0	0	$1.6 \cdot 10^{-10}$
9	0	0	0	0	0	0	$5.4 \cdot 10^{-13}$
Total	10000	10000	10000	10000	10000	10000	10000
Pear	17.25	3.45	17.12	88539	6.14	7.57	

Table: Numbers of segments resulting in k (out of 11) p-values smaller than 0.01 and values of Pearson statistics for truncated binomial distribution

After exclusion of one variant of Serial test and both variants of Cumulative Sums test the correspondence with the hypothesis on independence became better, but nevertheless is far from normal.

If χ^2_2 denotes random variable having the chi-square distribution with 2 degrees of freedom (i. e. exponential distribution with parameter $\frac{1}{2}$ and mean 2), then $\mathbf{P}\{\chi^2_2 \geqslant 17.12\} \approx 0.0002$, $\mathbf{P}\{\chi^2_2 \geqslant 3.45\} \approx 0.178$, $\mathbf{P}\{\chi^2_2 \geqslant 6.14\} \approx 0.046$, $\mathbf{P}\{\chi^2_2 \geqslant 7.57\} \approx 0.023$.

Note also that the sequence AES_0 was obtained with zero key, sequence MixL_1 was constructed from sequences having characteristic polynomials of the same degree 33, sequence MixL_2 contains segments of recurrent sequence with three-term characteristic polynomial.

Testing independence of 162 tests

k	AES_1	MixL_1	MixL_2	$\mathbf{E}\eta_k$
0	2176	2178	1450	1962
1	3106	3019	1401	3212
2	2390	2481	1724	2611
3	1323	1353	1477	1407
4	597	594	1191	564
5	275	259	838	180
6	94	83	544	47.7
7	27	25	361	10.7
8	8	5	237	2.1
9	3	3	159	0.36
10	0	0	94	0.06
11	1	0	75	0.008
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
141	0	0	1	$1.03 \cdot 10^{-252}$
Total	10000	10000	10000	10000
$Pear$	197	135	149600	

Table: Numbers of segments rejecting by k (out of 162) tests

Conclusion

A new approach to testing independence of statistical tests based on frequencies of segments simultaneously rejected by given number of tests is suggested and applied to the NIST package of tests.

The hypothesis on the independence of tests included in the NIST Statistical Test Suite was rejected.

It is shown that after exclusion of some tests from 15 tests of NIST package remaining tests look more independent than tests of the whole package.

It is shown that pseudorandom sequences constructed by alternating concatenation of segments of two linear recurrent sequences over GF(2) with non-regularly varying lengths do not rejected by the NIST package if these recurrent sequences are not too simple.

Thank you for attention!