On the correlation and sensitivity of so far statistical randomness tests based on runs

Hoang Dinh Linh, Do Dai Chi, Nguyen Tuan Anh, Le Thao Uyen

***Abstract*—Random numbers play a very important role in cryptography. More precisely, almost cryptographic primitives are ensured their security based on random values such as random key, nonces, salts... Therefore, the assessment of randomness according to statistical tests is really essential for measuring the security of cryptographic algorithms. In this paper, we focus on so far randomness tests based on runs in the literature. First, we have proved in detail that the expected number of gaps (or blocks) of length i in a random sequence of length n is $e_i = (n - i + 3)/2^{i+2}$. Secondly, we have evaluated correlation of some tests based on runs so far using Pearson coefficient method [5, 6] and Fail-Fail ratio one [7, 8]. Surprisingly, the Pearson coefficient method do not show any strong linear correlation of these runs-based tests but the Fail-Fail ratio do. Then, we have considered the sensitivity of these runs tests with some basic transformations. Finally, we have proposed some new runs tests based on the sensitivity results and applied evaluations to some random sources.**

***Tóm tắt*—Số ngẫu nhiên đóng một vai trò quan trọng trong mật mã. Cụ thể, độ an toàn của hầu hết các nguyên thủy mật mã đều được đảm bảo dựa trên các giá trị ngẫu nhiên như khóa, nonce, salt… Do đó, việc đánh giá tính ngẫu nhiên dựa trên các kiểm tra thống kê là thực sự cần thiết để đo độ an toàn cho các thuật toán mật mã. Trong bài báo này, chúng tôi tập trung vào các kiểm tra ngẫu nhiên dựa vào run trong các tài liệu. Đầu tiên, chúng tôi chứng minh chi tiết rằng kỳ vọng số các gap (khối) độ dài i trong một chuỗi ngẫu nhiên độ dài n là $e_i = (n - i + 3)/2^{i+2}$. Sau đó, chúng tôi đánh giá mối tương quan của một số kiểm tra dựa vào run bằng phương pháp hệ số Pearson [5, 6] và tỷ số Fail-Fail [7, 8]. Đáng ngạc nhiên là phương pháp hệ số Pearson không cho thấy bất kỳ mối tương quan tuyến tính mạnh nào của các kiểm tra dựa vào run, trong khi đó tỷ số Fail-Fail lại chỉ ra. Tiếp theo, chúng tôi xem xét độ nhạy của các kiểm tra run này với một số phép biến đổi cơ bản. Cuối cùng, chúng tôi đề xuất một số kiểm tra run mới dựa trên các kết quả độ nhạy và đánh giá áp dụng chúng cho một số nguồn ngẫu nhiên.**

***Keywords*—Randomness testing, Runs, Correlation, Sensitivity.**

***Từ khóa*—Kiểm tra ngẫu nhiên, Runs, tương quan, độ nhạy.**

I. INTRODUCTION

Statistical randomness tests play an important role in assessing the security of cryptographic algorithms. There have been many independently statistical randomness tests in the literature. Knuth [1] presented a number of statistical tests including frequency check, serial test, poker test, series test (run) ... Another test suite is the DIEHARD tests [2] including 18 statistical tests. In addition, there is a Crypt-XS test suite [3] proposed by the Information Security Research Center of Queensland University of Technology. Finally, the currently widely used test suite is the SP 800-22 statistical test suite [4] originally developed by NIST with 16 tests but then shortened to 15 tests (omitted Lempel-Ziv complexity test).

In addition, there are a number of randomness testing standards that are not presented in test suites or independently used. In 1992 Maurer proposed a universal statistical test for random bit generators. In 2004, Hernandez et al. proposed a new test called the Strict Avalanche Criterion (SAC) ... And recently, Doğanaksoy et al. [5] proposed three new randomness tests based on the length of runs in 2015.

Our Contributions. In this paper, we present some results on correlation and sensitivity of randomness tests based on runs so far. Specifically, we have shown that the expected number of gaps (or blocks) of length *i* in a binary random sequence of length *n* should be $\frac{n-i+3}{2^{i+2}}$.

This manuscript is received on October 29, 2021. It is commented on October 29, 2021 and accepted on November 10, 2021 by the first reviewer. It is commented on October 29, 2021 and accepted on November 15, 2021 by the second reviewer.

Furthermore, we have evaluted the correlation and sensitivity of some runs tests in the literature. Finally, we have proposed some new runs tests based on the sensitivity results and applied evaluations to some random sources.

Construction. The rest of the paper includes: Section 2 presents three postulates on randomness given by Golomb [6] as well as some tests based on runs, the Pearson coefficient method, Fail-Fail ratio and some basic transformations. Some results of correlation of these run test are presented in section 3. In section 4, we present the result of sensitivity of these runs tests with some transformations and propsed some new runs tests. Finally, the conclusions and future research directions are presented in Section 5.

II. PRELIMIARIES

In this section, we present the three postulates of randomness given by Golomb in [6]. This is one of the bases for assessing the randomness of a sequence. Then, we outlined some statistical tests related to the runs as well as the reasons for studying new statistical tests based on length of runs.

A. Golomb's Randomness Postulates

Let $s = s_0, s_1, \cdots, s_{n-1}, \cdots$ be an infinite binary sequence periodic with *n* or a finite sequence of length *n*. A run is defined as an uninterrupted maximal sequence of identical bits. Runs of 0's are called *gap*; runs of 1's are called *block*. R1, R2, and R3 are Golomb's randomness postulates which are given as follows:

(R1) In a period of *s*, the number of 1's should differ from the number of 0's by at most 1. In other words, the sequence should be balanced.

(R2) In a period of *s*, at least half of the total number of runs of 0's or 1's should have length one, at least onefourth should have length 2, at least one-eighth should have length 3, and so on. Moreover, for each of these lengths, there should be (almost) equally many gaps and blocks.

(R3) The autocorrelation function $C\ t$ should be two-valued. That is, for some integer *K* and for all $t = 0,1,\cdots,n-1$,

$$C\ t\ = \sum_{i=0}^{n-1} -1^{\ s_i + s_{i+t}} = \begin{cases} n, t = 0 \\ K, 1 \le t \le n-1. \end{cases}$$

In this paper, we mainly focus on the first and second postulates, and the last one is not a matter of concern.

B. Some basic run tests

Golomb's second postulate is on the number of runs in a sequence. Tests which consider the number of runs, are called run tests and are also included in many test suites such as Knuth [1], DIEHARD [2], TestU01 [7], NIST [4], Handbook of Applied Cryptography (abbreviated as Handbook) [8]. Since calculating the expected number of fixed-length runs in a random sequence is a difficult task (especially when the length of runs is large), most tests only consider the total number of runs and do not consider the number of runs of different lengths as follows:

1. *Run tests in Knuth [1] and DIEHARD [2] test suites*

These test suites define the run test on random numbers. They define runs as *runs up* and *runs down* in a sequence. For example, consider a sequence of length 10, $s_{10} = 138742975349$. Runs are indicated by putting a vertical line between s_j 's when $s_j > s_{j+1}$. Therefore, runs of the sequence 138742975349 can be seen as $|138|7|4|29|7|5|349|$. In other words, the run test examines the length of monotone subsequences.

2. *Run test in TestU01 [7] test suite*

This test suite defines runs and gap tests for checking the randomness of long binary sequences of length *n*. This test calculates the runs of 1 and 0 until the total number of runs is $2r$. Next, number of runs 1 and 0 correspond to each length of $j = 1,2,\cdots,k$ is calculated and recorded. Finally, applying χ^2 test on these values. Test of the longest run of one also be

defined for subsequences of length *m* that obtained from the original binary sequence of length *n*.

3. Run test in NIST [4] test suite

NIST test suite is widely used to checks the randomness of pseudorandom sequences. In the suite, 2 of 15 tests are variations of run tests. They are called run test and longest run of ones in a block test. The first one deals with the total number of runs in a sequence. It calculates the total number of runs in a sequence and determines whether it is consistent with the expected number of runs, which is supposed to be close to $n/2$ in a sequence or not. The second one determines whether the longest run of ones in the sequence is consistent with the length of the longest runs of ones which is in a random sequence. In NIST test suite the reference distributions for the run tests are a χ^2 distribution.

4. Test based on the length of runs [5]

A. Doğanaksoy et al. [5] have proposed three new randomness tests based on length of run in 2015. These tests consider runs of length one, runs of length two and runs of length three of subsequences of length *m*. Finally, applying χ^2 test on these values. In [9], H.D.Linh have shown that some probability values given in [5] are inaccurate, and this may lead to a mistake in assessing the randomness of the input sequences, thereby giving to incorrect assertions about the security of cryptographic algorithms.

5. Runs test in Handbook [8]

There are five basic tests in this book [8]. The purpose of the runs test is to determine whether the number of runs of various lengths in the sequence s is as expected for a random sequence. The expected number of gaps (or blocks) of length *i* in a random sequence of length *n* is $e_i = \frac{n-i+3}{2^{i+2}}$ with no proof. Let *k* be equal to the largest integer *i* for which $e_i \geq 5$. Let B_i, G_i be the number of blocks and gaps, respectively, of length *i* in *s* for each $i, 1 \leq i \leq k$. The reference distributions for the runs tests are a χ^2 distribution. In this paper, we have proof in detail that the expected number of gaps (or blocks) of length *i* in a random sequence of length *n* is $e_i = \frac{n-i+3}{2^{i+2}}$.

C. The Pearson coefficients

To evaluate the correlation of tests, we use the Pearson's correlation coefficient method [5, 6] which measures a linear relation and allows the detection of linear correlations between tests based on their statistical values.

The Pearson's correlation coefficient is the test statistics that measures the strength and direction of a linear relationship between two random variables.

Definition 1. (**Pearson's correlation coefficient**) Let *X* and *Y* be two random variables. The Pearson's correlation coefficient between two random variables *X* and *Y*, denoted as r_{XY}, is given by the following equation:

$$r_{XY} = Corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y},$$

where $cov(X,Y)$ is the covariance between *X* and *Y*, and σ_X, σ_Y are, respectively, the standard deviation of *X* and *Y*.

Correlation coefficient interpretion: The Pearson correlation value r_{XY} can take on any value in the range [-1, 1], where

- $r_{XY} = 0$: indicates that there is no linear relationship between the variables X and Y. However, it is possible that the variables have a non-linear relationship.
- $r_{XY} \in \{1, -1\}$: implies that a linear equation describes the relationship between X and Y perfectly.
- $r_{XY} < 0$: *signifies a negative correlation*, that is both variables X, Y move in the opposite direction (as the value of one variable increases, the value of the other variable decreases, and vice versa).
- $r_{XY} > 0$: *indicates a positive correlation*, that is both variables X, Y move in the same direction (as the value of one variable

to increase or decrease together with another variable).

- If $r_{XY} \in (-1,-0.5) \cup (0.5,1)$, then X and Y said to be a strong correlation.
- If $r_{XY} \in (-0.49,-0.3) \cup (0.3,0.49)$, then X and Y have a moderate correlation.
- If $r_{XY} \in (-0.29,0) \cup (0,0.29)$, then X and Y have a weak correlation.

D. Fail-Fail ratio

Another measure used to evaluate the correlation between randomness tests is the Fail-Fail ratio [7, 8]. As we know, a statistical test for randomness will determine whether the sequence could be passed the test, i.e whether the sequence passed or failed in the test. Therefore, it is advisable to base the correlation on the failed sequences. The behaviour of a test on failed sequences from another can give some idea of the correlation between these two tests. For example, if the tests T_1 and T_2 are uncorrelated, then it is expected that the probability that a sequence has failed for T_1 will fail for T_2 is equal to the fail probability of any sequence in the sequence space. That is, a failing from test T_1 should not affect the result of the test T_2. Any other results can indicate a relationship between the tests mentioned above. In such manner, the fail ratio of both T_1 and T_2 on the sequences to fail from T_1 or from T_2 can be used as a correlation measure. We call this probability *Fail-Fail Ratio* (FFR). To do this, we evaluate the following:

Step 1: Fixed two tests T_i, T_j and $\boldsymbol{S}$ denotes the set of sample sequences. Let $F_{T_i} = \{s \in \boldsymbol{S} | T_i(s) < 0.01\}$ be the set of all the failed sequences for the test T_i and $F_{T_j} = \{s \in \boldsymbol{S} | T_j(s) < 0.01\}$ be the set of all the failed sequences for test T_j. Correlation between T_i and T_j is the ratio of the sequences in F_{T_i} also fails from T_j and given by

$$Cor(T_i, T_j) = \frac{\left|F_{T_i} \cap F_{T_j}\right|}{\left|F_{T_i}\right|}.$$

Similarly, we have the correlation between T_j and T_i defined by

$$Cor(T_j, T_i) = \frac{\left|F_{T_i} \cap F_{T_j}\right|}{\left|F_{T_j}\right|}.$$

Step 2: After calculating the above correlation values for all pairs (T_i, T_j), we will build a fail-fail ratio table to analyze the correlation between the tests. Note that fail-fail ratio table is not symmetric since such operations are not commutative.

Step 3: If $Cor(T_i, T_j) > 0.05$ and $Cor(T_j, T_i) > 0.05$, then thì we can consider T_i and T_j correlate with each other.

E. Transfomations and Sensitivity

We will let $s = s_0 s_1 \dots s_{n-1}$ to denote a binary sequence of length n. We present some of the basic transformations can be used to evaluate and analyze the sensitivity of statistical tests.

Complement: This transformation is applying the bitwise NOT operator to the sequence: $C(s) = \tilde{s}$ such that $\hat{s}_i = 1 \oplus s_i$, where $\tilde{s} = (\hat{s}_0, \dots, \hat{s}_{n-1})$.

Reverse: This transformation reverses the sequence: $R(\boldsymbol{s}) = s_{n-1} s_{n-2} \dots s_0$, where $s = s_0 \dots s_{n-2} s_{n-1}$.

Swap Bits: This transformation swaps two successive bits in the sequence: $B(s) = \tilde{s}$ such that $\tilde{s} = s_2 s_1 s_4 s_3 \dots$

Swap Halves: Swaps the first half of the sequence with the last half: if $s = A \| B$, then $H(s) = B \| A$.

Swap Half-Reverse Last: first swaps the halves of the sequence and then reverses the last half of the swapped sequence: if $s = A \| B$ then $HR(s) = B \| R(A)$.

Reverse Halves: This transformation reverses each half of the sequence: if $s = A \| B$, then $RH(s) = R(A) \| R(B)$.

Complement Reverse: This transformation reverses the complement of the sequence $CR(s) = C(R(s))$.

Definition 2 (Invariant, [10]). Let T be a statistical randomness test and $\pi : L \to L$ be a transformation, where L is a set of all binary sequence of length n bits. Then, T is called *invariant* under π if with any $S \in L$, we have $T(S) = T(\pi(S))$.

We define the concept of *sensitivity* to statistical randomness testing as follows.

Definition 3 (Sensitivity, [10]). Let T be a statistical randomness test and $\pi : L \to L$ be a transformation, where L is a set of all binary sequence of length n bits. The sensitivity of test T under transformation π is defined as a measure of the effect of a transformation, denoted by $\text{sen}(T_\pi)$. where

- If T is invariant under π, then the sensitivity of T under π is defined by 0, denoted as $\text{sen}(T_\pi) = 0$.
- If the transformation π has a small effect on the test results (that is, there is a significant correlation between $T(S)$ and $T(\pi(S))$), then the sensitivity is defined as 1, denoted $\text{sen}(T_\pi) = 1$.
- If $T(S)$ and $T(\pi(S))$ are statistically independent, then the sensitivity is given by 2, denoted $\text{sen}(T_\pi) = 2$. In this case, $T(\pi(\cdot))$ can be added to the statistical test suites as a new test.

One method of observing the correlation between two randomness tests is to analyze the response of tests to changes in the sequence. Applying basic transformations to input sequences that significantly changes the output p-values of a randomness test as an alternative to developing more tests. Specifically:

Step 1: We first use the test T to a set of sample sequences; then also continue to the new sequences which is the output of applying transformation π to the sample sequences.

Step 2: We will calculate the correlation value between $T(\mathbf{s})$ and $T(\pi(\mathbf{s}))$ by using the "*Pearson correlation coefficient*" (see Section 2.3), where $\mathbf{s}$ is the original sequence, $\pi(\mathbf{s})$ is the transformed sequence. From the obtained Pearson's correlation value, we give an evaluation to sensitivity of the test T under transformation π.

Step 3: If the correlation value geater than 0.80, then shows a correlation of test T under transformation π.

III. CORRELATION OF SOME TESTS BASED ON RUNS

In this paper, we will consider the independence between 6 runs-based tests include: NIST's runs test, longest run of ones in a block of NIST, 3 runs tests based on length 1 , 2, 3 by A. Doğanaksoy et al. and runs test in Handbook. In these runs tests, runs tests in Handbook are probably general tests and most closely related to Golomb's second postulate. However, in [8] only provided a description of this test without detailed proof of the distribution of the statistics mentioned. In this paper, we provided a detailed proof that the expected number of gaps (or blocks) of length i in a random sequence of length n is $e_i = \frac{n-i+3}{2^{i+2}}$.

A. Distribution of runs test in Handbook

In a random sequence of length n, [8] claim that the expected number of gaps (or blocks) of length i is $e_i = (n - i + 3)/2^{i+2}$. Let k is the largest integer i for which $e_i \geq 5$ and B_i, G_i be the number of blocks and gaps, respectively, of length i in s for each i, $1 \leq i \leq k$. The statistic used is

$$X_4 = \sum_{i=1}^{k} \frac{(B_i - e_i)^2}{e_i} + \sum_{i=1}^{k} \frac{(G_i - e_i)^2}{e_i}$$

According [8], this statistic approximately follows a χ^2 distribution with $2k - 2$ degrees of freedom. The above arguments are given by [8] without any proof.

In this section, we will prove that $e_i = (n - i + 3)/2^{i+2}$. We refer to the paper [11] to clarify this argument. Mood [11] gave the

distribution of runs with the general formula. However, Mood's proof is quite complicated to get e_i. This paper provides an easier way to get it.

Fact 1. *The expected number of gaps (or blocks) of length i in a random sequence of length n is $e_i = (n-i+3)/2^{i+2}$.*

Proof. Consider a random sequence of length n: $s = s_0 \dots s_{n-1}$, let n_0 denote the number of bit of 0 and let n_1 denote the number of bit of 1, $n = n_0 + n_1$. Let r_{1i} denote the number of blocks of length i, and let r_{0i} denote the number of gaps of length i. For example, we consider the following sequence

$$s = 0110100011000$$

We have, $r_{01} = 2, r_{03} = 2, r_{11} = 1, r_{12} = 2$ and all other $r_{i,j} = 0$.

Our problem is computing $E(r_{1i})$ because of $E(r_{1i}) = E(r_{0i})$. We let $r_1 = \sum_i r_{1i}$ and $r_0 = \sum_i r_{0i}$ denote the total number of blocks and gaps respectively. In this paper, a binomial coefficient will be denoted by

$$\binom{m}{k} = \frac{m!}{k!\,(m-k)!}, \tag{1}$$

And this is defined to be zero when $m < k$. A multinomial coefficient will often be denoted by

$$\begin{bmatrix} m \\ m_1, m_2, \dots, m_i \end{bmatrix} = \frac{m!}{m_1!\, m_2! \dots m_l!} \tag{2}$$

$$\sum m_i = m, \;\; m_i \geq 0 \tag{3}$$

and when such a coefficient is to be summed over the indices m_i the condition (3) are always understood and will not be repeated, other conditions on the indices will be places below the summation sign.

(The formula (2) is the coefficient of the following polynomial with m, s are positive integers

$$(x_1 + x_2 + \dots + x_l)^m = \sum_{m_1+\dots+m_l=m} \begin{bmatrix} m \\ m_1, m_2, \dots, m_l \end{bmatrix} \prod_{i=1}^{l} x_t^{m_i})$$

For simplicity, we can write $\begin{bmatrix} m \\ m_i \end{bmatrix}$ instead of $\begin{bmatrix} m \\ m_1, m_2, \dots, m_l \end{bmatrix}$.

Firstly, we consider the fix for the case n_0 and n_1 are fixed. Given a set of numbers r_{ij} (i.e, set of number of blocks or gaps of length j) $(i = 0,1; j = 1, 2, \dots, n_i)$ such that $\sum_j j r_{ij} = n_i$, there are $\begin{bmatrix} r_1 \\ r_{1j} \end{bmatrix}$ and $\begin{bmatrix} r_0 \\ r_{0j} \end{bmatrix}$ different arrangements of blocks and gaps. Hence the total number of ways of obtaining the set r_{ij} is

$$N(r_{ij}) = \begin{bmatrix} r_1 \\ r_{1j} \end{bmatrix} \begin{bmatrix} r_0 \\ r_{0j} \end{bmatrix} F(r_1, r_0), \tag{4}$$

where $F(r_1, r_0)$ is the number of ways of arranging r_1 blocks and r_0 gaps such that no two adjacent blocks or gaps are the same kind. Thus

$$\begin{aligned} F(r_1, r_0) \;\; &= 0 \text{ if } |r_1 - r_0| > 1 \\ &= 1 \text{ if } |r_1 - r_0| = 1 \\ &= 2 \text{ if } r_1 = r_0 \end{aligned} \tag{5}$$

Since there are $\binom{n}{n_1}$ possible arrangements of s (note that, we consider that fixed n_1, n_0 here), we have the distribution of the r_{ij}

$$P(r_{ij}) = \frac{\begin{bmatrix} r_1 \\ r_{1j} \end{bmatrix} \begin{bmatrix} r_0 \\ r_{0j} \end{bmatrix} F(r_1, r_0)}{\binom{n}{n_1}}. \tag{6}$$

We will compute the distribution of r_{1j}. We sum $\begin{bmatrix} r_0 \\ r_{0j} \end{bmatrix}$ over all partitions of n_0. This is accomplished by finding the coefficient of x^{n_0} in

$$(x + x^2 + x^3 + \cdots)^{r_0}$$

$$= x^{r_0}(1 + x + x^2 + \cdots)^{r_0} = \frac{x^{r_0}}{(1-x)^{r_0}}$$

$$= x^{r_0} \sum_{t=0}^{\infty} \binom{r_0 - 1 + t}{r_0 - 1} x^t.$$

(Using Taylor expansion of $\frac{1}{(1-x)^{r_0}}$.)

We consider the coefficient of x^{n_0} in the right hand. For $t = n_0 - r_0$, we have the coefficient of x^{n_0} is: $\binom{n_0 - 1}{r_0 - 1}$.

Next, we consider the coefficient of x^{n_0} in the left hand. We have

$$(x + x^2 + x^3 + \cdots)^{r_0}$$

$$= \sum_{r_{01}+r_{02}+\dots=r_0} \begin{bmatrix} r_0 \\ r_{01}, r_{02}, \dots \end{bmatrix} \prod_{j=1} (x^j)^{r_{0j}}$$

$= \sum_{r_{01}+r_{02}+\cdots=r_0} \begin{bmatrix} r_0 \\ r_{01}, r_{02}, \ldots \end{bmatrix} \prod_{j=1} x^{jr_{0j}}$

The coefficient of x^{n_0} is:

$$\sum_{\sum jr_{0j}=n_0} \begin{bmatrix} r_0 \\ r_{0j} \end{bmatrix}.$$

Thus, we have

$$\sum_{\sum jr_{0j}=n_0} \begin{bmatrix} r_0 \\ r_{0j} \end{bmatrix} = \binom{n_0 - 1}{r_0 - 1}. \quad (7)$$

We have then

$$P(r_{1j}, r_0) = \frac{\begin{bmatrix} r_1 \\ r_{1j} \end{bmatrix} \binom{n_0 - 1}{r_0 - 1} F(r_0, r_1)}{\binom{n}{n_1}} \quad (8)$$

and summing this over r_0 we have

$$P(r_{1j}) = \sum_{r_0} \frac{\begin{bmatrix} r_1 \\ r_{1j} \end{bmatrix} \binom{n_0 - 1}{r_0 - 1} F(r_0, r_1)}{\binom{n}{n_1}}$$

$$= \frac{\begin{bmatrix} r_1 \\ r_{1j} \end{bmatrix} \binom{n_0 - 1}{r_1 - 2}}{\binom{n}{n_1}} + \frac{2\begin{bmatrix} r_1 \\ r_{1j} \end{bmatrix} \binom{n_0 - 1}{r_1 - 1}}{\binom{n}{n_1}}$$

$$+ \frac{\begin{bmatrix} r_1 \\ r_{1j} \end{bmatrix} \binom{n_0 - 1}{r_1}}{\binom{n}{n_1}} = \frac{\begin{bmatrix} r_1 \\ r_{1j} \end{bmatrix} \binom{n_0 + 1}{r_1}}{\binom{n}{n_1}}. \quad (9)$$

Denote

$$x^{(a)} = x(x-1)(x-2)\ldots(x-a+1) = \frac{x!}{(x-a)!}.$$

We have:

$$\sum_{(1)} \begin{bmatrix} r_1 \\ r_{1i} \end{bmatrix} \prod_i r_{1i}^{(a_i)} = r_1^{(\sum a_i)} \binom{n_1 - \sum ia_i - 1}{r_1 - \sum a_i - 1}, \quad (10)$$

where $\sum_{(1)}$ denotes summation over all positive integers $r_{11}, r_{12}, \ldots, r_{1n_1}$ such that $\sum_1^{n_1} ir_{1i} = n_1$. The formula (10) can be verified by differentiating

$$\psi(t_i) = (t_1 x + t_2 x^2 + \cdots)^{r_1}$$

a_i times with respect to t_i $(i = 1,2,\ldots,n_1)$, then finding the coefficient of x^{n_1} after putting $t_i = 1$. Indeed, we have

$$\psi(t_i) = (t_1 x + t_2 x^2 + \cdots)^{r_1}$$

$$= \sum_{r_{11}+r_{12}+\cdots=r_1} \begin{bmatrix} r_1 \\ r_{11}, r_{12}, \ldots \end{bmatrix} \prod_{j=1} \left(t_j x^j\right)^{r_{1j}}$$

$$= \sum_{r_{11}+r_{12}+\cdots=r_1} \begin{bmatrix} r_1 \\ r_{11}, r_{12}, \ldots \end{bmatrix} \prod_{j=1} t_j^{r_{1j}} x^{jr_{1j}}.$$

We take the derivative $\psi(t_i)$ a_1 times with respect to t_1:

$$\psi_{a_1}(t_i) = \sum_{r_{11}, r_{12}, \ldots} \begin{bmatrix} r_1 \\ r_{11}, r_{12}, \ldots \end{bmatrix} r_{11}^{(a_1)} x^{1r_{11}} t_1^{r_{11}-a_1} \prod_{j=2} t_j^{r_{1j}} x^{jr_{1j}}.$$

We take the derivative $\psi_{a_1}(t_i)$ a_2 times with respect to t_2:

$$\psi_{a_1 a_2}(t_i) = \sum_{r_{11}, r_{12}, \ldots} r_{11}^{(a_1)} x^{1r_{11}} t_1^{r_{11}-a_1} r_{12}^{(a_2)} x^{2r_{12}} t_2^{r_{12}-a_2} \cdot \prod_{j=3} t_j^{r_{1j}} x^{jr_{1j}}.$$

Then we have:

$$\psi_{a_1 \ldots a_{n_1}} = \sum_{r_{11}, r_{12}, \ldots} \begin{bmatrix} r_1 \\ r_{11}, r_{12}, \ldots \end{bmatrix} \prod_{j=1}^{n_1} r_{1j}^{(a_j)} x^{jr_{1j}} t_j^{r_{1j}-a_j} \prod_{j=n_1} t_j^{r_{1j}} x^{jr_{1j}}.$$

Put $t_i = 1$, the coefficient of x^{n_1} is

$$\sum_{r_{11}, r_{12}, \ldots, r_{1n_1}} \begin{bmatrix} r_1 \\ r_{11}, r_{12}, \ldots, r_{1n_1} \end{bmatrix} \prod_{j=1}^{n_1} r_{1j}^{(a_j)}.$$

In the other hand, we can take the derivative $\psi(t_i)$ by the following way. We take the derivative $\psi(t_i)$ a_1 times with respect to t_1.

$$\psi_{a_1}(t_i) = r_1^{(a_1)} x^{1a_1} (t_1 x + t_2 x^2 + \cdots)^{r_1 - a_1}.$$

Then, we take the derivative $\psi_{a_1}(t_i)$ a_2 times with respect to t_2.

$$\psi_{a_1 a_2}(t_i) = r_1^{(a_1+a_2)} x^{1a_1+2a_2} \cdot (t_1 x + t_2 x^2 + \cdots)^{r_1 - a_1 - a_2}.$$

We have then

$$\psi_{a_1 a_2 \ldots a_{n_1}}(t_i)$$

$$= r_1^{(\sum_1^{n_1} a_i)} x^{\sum_1^{n_1} j a_j (t_1 x + t_2 x^2 + \cdots)^{r_1 - \sum_1^{n_1} a_j}}$$

$$= r_1^{(\sum_1^{n_1} a_i)} x^{r_1 - \sum_1^{n_1} a_j + \sum_1^{n_1} j a_j}$$

$$\cdot (t_1 + t_2 x + \cdots)^{r_1 - \sum_1^{n_1} a_j}$$

Put $t_i = 1$ we have

$$\psi_{a_1 a_2 \ldots a_{n_1}}(1) = \frac{r_1^{(\sum_1^{n_1} a_i)} x^{r_1 - \sum_1^{n_1} a_j + \sum_1^{n_1} j a_j}}{(1-x)^{r_1 - \sum_1^{n_1} a_j}}$$

$$= r_1^{(\sum_1^{n_1} a_i)} x^{r_1 - \sum_1^{n_1} a_j + \sum_1^{n_1} j a_j}$$

$$\cdot \sum_{t=0}^{\infty} \binom{r_1 - \sum_1^{n_1} a_j - 1 + t}{r_1 - \sum_1^{n_1} a_j - 1} x^t.$$

We compute the coefficient of x^{n_1} in this expression. We have $t = n_1 - r_1 - \sum_{j=1}^{n_1} j a_j + \sum_{j=1}^{n_1} a_j$, and the coefficient of x^{n_1} is:

$$r_1^{(\sum_1^{n_1} a_i)} \binom{n_1 - \sum_1^{n_1} j a_j - 1}{r_1 - \sum_1^{n_1} a_j - 1}.$$

Compare two ways of computing the coefficient x^{n_1} we have (10).

Using the formula $E\big(f(r)\big) = \sum_r f(r) P(r)$, we have:

$$E\left(\prod_i r_{1i}^{(a_i)}\right)$$

$$= \sum_{r_{1i}} \prod r_{1i}^{(a_i)} \begin{bmatrix} r_1 \\ r_{1i} \end{bmatrix} \binom{n_0 + 1}{r_1} \Big/ \binom{n}{n_1}$$

$$= \sum r_1^{(\sum a_i)} \binom{n_1 - \sum i a_i - 1}{r_1 - \sum a_i - 1} \binom{n_0 + 1}{r_1} \Big/ \binom{n}{n_1}$$

$$= \sum (n_0 + 1)^{(\sum a_i)} \binom{n_1 - \sum i a_i - 1}{r_1 - \sum a_i - 1}$$

$$\cdot \binom{n_0 - \sum i a_i + 1}{r_1 - \sum i a_i} \Big/ \binom{n}{n_1} \qquad (11)$$

$$= (n_0 + 1)^{(\sum a_i)} \binom{n - \sum (i+1) a_i}{n_1 - \sum i a_i} \Big/ \binom{n}{n_1}.$$

The sum on r_1 involved in the last step is given by the identity

$$\sum_{i=0}^{B} \binom{A}{C+i} \binom{B}{i} = \binom{A+B}{C+B},$$

which is readily obtained by equating coefficients x^C in

$$(1+x)^A \left(1 + \frac{1}{x}\right)^B = \frac{(1+x)^{A+B}}{x^B}.$$

Now, instead of fixing the values of n_0 and n_1, we assume that they are selected randomly and have only one constraint $\sum n_i = n$. The probability of occurrence of bits 0 and 1 is p_0 and p_1, respectively. We have the fundamental relation

$$P(X, Y) = P(X|Y) P(Y) \qquad (12)$$

where, X will represent the set of variables r_{ij} or r_i, and Y the variables n_i. We have

$$P(n_1, n_0) = \binom{n}{n_1} p_1^{n_1} p_0^{n_0}$$

We have then

$$E\big(f(X) g(Y)\big) = \sum_{XY} f(X) g(Y) P(X, Y)$$

$$= \sum_Y g(Y) P(Y) \left[\sum_X f(X) P(X|Y)\right] \qquad (13)$$

Using (9), (10) and (13) we have

$$E\left(n_1^{(a)} \prod_1^{n_1} r_{1i}^{(a_i)}\right)$$

$$= \sum_{n_1=0}^{n} n_1^{(a)} \binom{n}{n_1} p_1^{n_1} p_0^{n_0}$$

$$\left[\sum_{r_{1i}} \prod_{i=1}^{n_1} r_{1i}^{(a_i)} \frac{\begin{bmatrix} r_1 \\ r_{1j} \end{bmatrix} \binom{n_0+1}{r_1}}{\binom{n}{n_1}}\right]$$

$$= \sum_{n_1}^{n} n_1^{(a)} \binom{n}{n_1} p_1^{n_1} p_0^{n_0}$$

$$\cdot \left[(n_0 + 1)^{(\sum a_i)} \binom{n - \sum (i+1) a_i}{n_1 - \sum i a_i} \Big/ \binom{n}{n_1}\right]$$

$$= \sum_{n_1=0}^{n} n_1^{(a)} (n_0 + 1)^{(\sum a_i)}$$

$$\cdot \binom{n - \sum i a_i - \sum a_i}{n_1 - \sum i a_i} p_1^{n_1} p_0^{n_0}$$

For $a = 0, a_i = 1$, and $a_j = 0$ with $j \neq i$, we have

$$E(r_{1i}) = \sum_{n_1=0}^{n} (n_0 + 1)^{(1)} \binom{n - i - 1}{n_1 - i} p_1^{n_1} p_0^{n_0}$$

$$= \sum_{n_1=0}^{n} [(n - i + 1) - (n_1 - i)]$$

$$\cdot \binom{n - i - 1}{n_1 - i} p_1^{n_1} p_0^{n_0}$$

$$= \sum_{n_1=0}^{n} \left[(n-i+1) \binom{n-i-1}{n_1-i} p_1^{n_1-i} p_0^{n_0-1} - (n_1-i) \binom{n-i-1}{n_1-i} p_1^{n_1-i} p_0^{n_0-1} \right] p_1^i p_0$$

$$= p_1^i p_0 \cdot$$

$$\left[(n-i+1) \sum_{n_1=0}^{n} \binom{n-i-1}{n_1-i} p_1^{n_1-i} p_0^{n_0-1} - (n-i-1) p_1 \sum_{n_1=0}^{n} \binom{n-i-2}{n_1-i-1} p_1^{n_1-i-1} p_0^{n_0-1} \right]$$

$$= p_1^i p_0 [(n-i+1) - (n-i-1) p_1].$$

Because s is the random sequence, we have $p_0 = p_1 = 1/2$. Thus:

$$E(r_{1i}) = \frac{n-i+3}{2^{i+2}}. \quad \blacksquare$$

B. Experimental Results

Since each randomness testing in different types of test suites will have distinct usage parameters, we selected the following parameters for the 6 run tests mentioned in this paper.

TABLE I. PARAMETERS USED FOR TEST BASED ON RUNS

Randomness Tests based on Runs	**Symbols**	**Parameters**
NIST Runs Test	T1	$\boldsymbol{n = 1000000}$
NIST Longest run of ones in a block	T2	$\boldsymbol{K = 6, M = 10000, n = 1000000}$
Run Test based on length 1	T3	$\boldsymbol{M = 512, n = 1000000, l = 1}$
Run Test based on length 2	T4	$\boldsymbol{M = 512, n = 1000000, l = 2}$
Run Test based on length 3	T5	$\boldsymbol{M = 512, n = 1000000, l = 3}$
Run Test in Handbook	T6	$\boldsymbol{n = 1000000}$

(M – block size, l – run length, K – number of probability interval, n – sequence length)

We use HMAC-DRBG generator with a true random input (random.org) to generate 10000 random sequences. Each sequence has 1000000 bit in length. For each run test, we evaluate 10000 p-values corresponding to these random sequences.

Using Pearson coefficient method, we have computed the Pearson coefficients as in Table II. Surprisingly, we do not see any strong linear correlation of these tests based on runs in this table. However, we found a weak correlation of T1 and T3 using Fail-Fail ratio method. Moreover, we found strong correlations of T6 and T1, T2, T3, T4 using the same one.

TABLE II. PEARSON CORRELATION VALUE OF RUN TESTS

	T1	**T2**	**T3**	**T4**	**T5**	**T6**
T1	**1.000**	0.013	0.277	0.060	0.010	0.182
T2	0.013	**1.000**	-0.011	-0.003	-0.009	0.148
T3	0.277	-0.011	**1.000**	0.005	-0.009	0.067
T4	0.060	-0.003	0.005	**1.000**	0.011	0.061
T5	0.010	-0.009	-0.009	0.011	**1.000**	0.069
T6	0.182	0.148	0.067	0.061	0.069	**1.000**

TABLE III. FAIL-FAIL RATIO OF RUN TESTS

	T1	T2	T3	T4	T5	T6
T1	**1.000000**	0.010526	**0.200000**	0.021053	0.010526	**0.905263**
T2	0.008696	**1.000000**	0.008696	0.008696	0.017391	**0.686957**
T3	**0.190000**	0.010000	**1.000000**	0.010000	0.040000	**0.500000**
T4	0.023529	0.011765	0.011765	**1.000000**	0.047059	**0.423529**
T5	0.009615	0.019231	0.038462	0.038462	**1.000000**	**0.519231**
T6	0.023344	0.021444	0.013572	0.009772	0.014658	**1.000000**

IV. SENSITIVITY OF SOME TESTS BASED ON RUNS

TABLE IV. THE SENSITIVITY OF RUNS-BASED TESTS UNDER TRANSFORMATION

	C(s)	B(s)	H(s)	R(s)	HR(s)	RH(s)	CR(s)
T1	1.000	**0.176**	1.000	1.000	1.000	1.000	1.000
T2	**0.007**	**0.264**	1.000	1.000	1.000	**0.007**	1.000

T3	1.000	1.000	1.000	1.000	1.000	1.000	1.000
T4	1.000	1.000	1.000	1.000	1.000	1.000	1.000
T5	1.000	1.000	1.000	1.000	1.000	1.000	1.000
T6	1.000	**0.248**	1.000	1.000	1.000	1.000	1.000

In order to analyse the sensitivity of runs-based tests under some basic transformations that we mentioned in above. We also use HMAC-DRBG generator with a true random input to generate 10000 random sequences. Each sequence has 1000000 bit in length. For each run test, we evaluate 10000 p-values corresponding to these random sequences. Then, we evaluate 10000 other p-values corresponding to transformed sequences for each transformation. Finally, we use the Pearson coefficient method to compute the Pearson coefficient of two sets of p-values. The results is shown in Table IV. From this result, we suggest 5 new tests based on runs: $T\_1\ (B(\cdot)), T\_2\ \big(C(\cdot)\big), T_2(B(\cdot)), T\_2\ (RH(\cdot))), T\_6\ (B(\cdot))$.

We have applied these test on some sample data (in NIST SP 800 - 22) and get the following results:

	data.pi	**data.e**	**data.sqrt2**	**data.sqrt3**
NIST Runs Test	0.419	0.562	0.313	0.261
NIST Longest run of ones in a block	0.024	0.719	0.012	0.447
Run Test based on length 1	0.037	0.538	0.570	0.850
Run Test based on length 2	0.629	0.886	0.419	0.919
Run Test based on length 3	0.421	0.660	0.782	0.269
Run Test in Handbook	0.001	0.130	0.029	0.283
New Runs Test 1: $T_1(B(\cdot))$	**0.521**	**0.249**	**0.966**	**0.249**
New Runs Test 2: $T_2(B(\cdot))$	**0.012**	**0.988**	**0.000**	**0.100**
New Runs Test 3: $T_6(B(\cdot))$	**0.016**	**0.008**	**0.001**	**0.010**
New Runs Test 4: $T_2(C(\cdot))$	**0.137**	**0.430**	**0.197**	**0.912**
New Runs Test 5: $T_2(RH(\cdot))$	**0.137**	**0.430**	**0.197**	**0.912**

Interestingly, using complement and reverse halves transformation before applying NIST Longest run of ones in a block will give the same result. So we have new test 4 and new test 5 are identical.

CONCLUSION

In this paper, we present some results on the correlation of runs-based tests and the sensitivity of these tests under basic transformations. We have proved in detail that the expected number of gaps (or blocks) of length i in a random sequence of length n is $e_i = (n - i + 3)/2^{i+2}$. Moreover, using Pearson coefficient method and Fail-Fail ratio method, we have evaluated correlation of some tests based on runs so far. Surprisingly, the Pearson coefficient method do not show any strong linear correlation of these runs-based tests but the Fail-Fail ratio do. Then, we have considered the sensitivity of these runs tests with some basic transformations. Finally, we have proposed some new runs tests based on the sensitivity results and applied evaluations to some random sources. However, in this paper, we still do not prove that the statistic X_4 has $2k - 2$ degrees of freedom. We will be complete it in the future.

REFERENCES

[1] MacLaren, M.D., The art of computer programming. Volume 2: Seminumerical algorithms (Donald E. Knuth). SIAM Review, 1970. 12(2): p. 306-308.

[2] Marsaglia, G., The marsaglia random number cdrom including the diehard battery of tests of randomness, 1995. 2008.

[3] Caelli, W., Crypt x package documentation. Information Security Research Centre School of Mathematics, Queensland University of Technology, 1992.

[4] Rukhin, A., et al., Statistical test suite for random and pseudorandom number generators for cryptographic applications, NIST special publication. 2010.

[5] Doğanaksoy, A., et al., New statistical randomness tests based on length of runs. Mathematical Problems in Engineering, 2015. 2015.

[6] Golomb, S.W., Shift register sequences. 1982: Aegean Park Press.

[7] L'Ecuyer, P. and R. Simard, TestU01: AC library for empirical testing of random number generators. ACM Transactions on Mathematical Software (TOMS), 2007. 33(4): p. 22.

[8] Menezes, A.J., P.C. Van Oorschot, and S.A. Vanstone, Handbook of applied cryptography. 2018: CRC press.

[9] Linh, H.Đ., Some results on new statistical randomness tests based on length of runs. Journal of Science Technology on Information security, 2018. 8(2): p. 10-18.

[10] Turan, M.S., A. DoĞanaksoy, and S. Boztaş. On independence and sensitivity of statistical randomness tests. in International Conference on Sequences and Their Applications. 2008. Springer.

[11] Mood, A.M., The distribution theory of runs. The Annals of Mathematical Statistics, 1940. 11(4): p. 367-392.

ABOUT THE AUTHORS

Hoang Đinh Linh

Workplace: Institute of Cryptographic Science and Technology, Vietnam Government Information Security Commission.

Email: hoangdinhlinh@bcy.gov.vn

Education: Bachelor of Mathatematic in VNU University of Science (2014).

Recent research direction: Private-key cryptograply, random statistical test, random number generator.

Do Dai Chi

Workplace: Institute of Cryptographic Science and Technology, Vietnam Government Information Security Commission.

Email: dodaichi2005@gmail.com

Education: Bachelor of Mathematic in VNU Univesity of Science (2013), Master of Cryptography in Limoges University (2019).

Recent research direction: Digital signature, provable security, public-key cryptography.

Nguyen Tuan Anh

Workpalce: Institute of Cryptographic Science and Technology, Vietnam Government Information Security Commission.

Email: tuananhnghixuan@gmail.com

Education: Bachelor of Mathematic in VNU University of Science (2016).

Recent research direction: Provable security, block cipher, message authentication code, hash function.

Le Thao Uyen

Workplace: Information Security Journal, Vietnam Government Information Security Commission

Email: ltuyen@bcy.gov.vn

Education: Bachelor of Engineering in Information Technology from Hanoi University of Science & Technology (2019).

Recent research direction: Cryptography, machine learning, deep learning, fully homomorphic encryption.