Analysis of Randomness of Runs and Its Application for Statistical Tests

Mohammad Dakhilalian[†] , Ebrahim molavian Jazi[†], Mohammad Jafar Taghiyar[†]

[†]Cryptography & System Security Research Laboratory
Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran

Summary
Statistical Tests are suitable means for analyzing properties of pseudorandom sequences, specifically in cryptography systems. Accordingly, various statistical tests have been proposed in literature. One of these tests is Runs Test. In this paper, we first state the common test for runs. Then by investigating the statistical behavior of runs in an *Ideal Random Sequence* (IRS), not only the runs test for total number of runs is improved but also two new tests are offered based on the distribution of runs of different lengths. The simulation results are also presented.
.

Key words:
Statistical Tests, Runs Test, Pseudorandom Sequence, Cryptography.

Introduction

Although, in many applications the use of random sequences is crucial, we cannot produce a random sequence in practice. In fact, the components of a random sequence are independent and identically distributed (i.i.d). Therefore random generators that produce sequences with properties closer to those of completely random sequences are of great interest. In this case, we need some tools to evaluate this closeness and thereby to evaluate how much a generator is random.

Since there are different definitions for randomness [15,16,18], there are also different methods and criteria to evaluate the randomness of a generator. The most common method for this purpose is the use of statistical tests. The essence of a statistical test is the mathematical analysis of a random sequence which is done by modeling one or more properties of the random sequence as a probabilistic model. Subsequently, these extracted properties are used to be compared to those of sequences under the test through a specific criterion.

Knowing the statistical characteristics of random sequences, some criteria have been proposed for pseudorandom sequences [4,5] and accordingly different tests have also been introduced for evaluating the randomness of sequences and their generators. The most common statistical tests have been proposed through different statistical test suites [6,8,9,10,11,12]. There is also a number of individual tests, see for example [13,14,17]. The most preferable method of statistically testing is through the use of Goodness-of-Fit theorem [7,9].

Statistical tests are carried out on a small portion of a sequence; for example on its first 1 million bits, through which it either fails or passes. To examine the randomness of a generator, usually a large number of its output sequences are tested and it is necessary that a specific number of sequences pass the test if it is about for the generator to be considered random. It must also be mentioned that success or failure of a sequence under a test does not signify whether the sequence is really random or not, since in statistical tests a typical behavior of completely random sequences is used for testing and only those sequences who agree with this typical behavior can pass the test.

Runs test is one of statistical tests which is used specially for assessing cryptographic algorithms and is discussed specifically in [6,10]. In this paper, the randomness of lengths of runs in a completely random sequence is modeled and analyzed thoroughly. Accordingly, an accurate method for calculating the exact mean and variance of the total number of runs and an improvement for statistic of Runs Test are presented [2]. Since a test which is based on evaluation of the probability distribution function itself is more suitable than a test which is based only on the mean and variance of the probability function, some chi-square tests based on the probability distribution function of runs, gaps and blocks are proposed in this paper. The simulation results on some well known test random sequences with empirical results on some generators are also presented.

Statistical Test on the Total Number of Runs

In this section we present an improved statistic for the test on the total number of runs presented in [5]. As discussed in the Introduction section, the aim of a statistical test is to compare a sequence with another one that has some specific properties. So we first present definition 1, in order to follow the rest of the paper more easily.

Manuscript received September 5, 2009
Manuscript revised September 20, 2009

Definition 1. An *Ideal Random Sequence (IRS)* is a sequence of independent and identically distributed random variables.

In an IRS, zeros and ones occur perfectly randomly without any order, thus appropriately modeling such sequences is of great importance. One of the properties of such sequences is the length and the number of runs of it.

Definition 2. An uninterrupted sequence of identical bits is called a *run*. If the run consists of ones then it is called a *block* and if it consists of zeros then it is called a *gap*.

For example the sequence $x^{18} = 010001100001111110$ comprises seven runs of lengths 1,2,3,4 and 6, three of them are blocks of lengths 1, 2 and 6 and the others are four gaps of lengths 1, 3 and 4.

In [5], the distribution of number of runs for an IRS $X^n = X_1, X_2, ..., X_n$ is computed as a Normal distribution with the following mean and variance (when $n \to \infty$):

$$\text{The mean of the number of runs } m = 1 + \frac{2n_0 n_1}{n}$$

$$\text{The variance of the number of runs } \sigma^2 = \frac{(m-1)(m-2)}{n-1} \qquad (1)$$

Where n_0, n_1 and n are the number of zeros, the number of ones and the length of the sequence under the test, respectively. Using (1), the test statistic for analyzing the statistical distribution of runs is presented in [5] as follows:

$$T_n(obs) = \frac{N_{obs} - m}{\sigma} \qquad (2)$$

In which N_{obs} is the total number of runs in the *n*-bit sequence. This test; however, is not accurate for statistical evaluation of the number of runs, since by using the statistical distribution of runs, one can introduce more exact tests as are presented in this paper.

In the next section, by investigating the randomness of runs in an IRS, we present an improved test for runs.

Idea of Test and Exact Mean and Variance

To clarify the idea of the test presented in this paper, we first explore a simple example.

Consider $x^{18} = x_1, x_2, ..., x_{18} = 010001100001111110$ in the previous example. As mentioned, this sequence has seven runs of lengths 1,2,3,4 and 6. If the sequence c^{17} is constructed from x^{18} as follows:

$$\begin{aligned} c^{17} &= c_1, c_2, ..., c_{17} \\ &= x_1 \oplus x_2, x_2 \oplus x_3, ..., x_{17} \oplus x_{18} \\ &= 11001010001000001 \end{aligned}$$

In which the sign $\oplus$ shows addition modulo 2 or XOR.

The c^{17} can also be considered as a sequence of some subsequences $c^{17} = r_1, r_2, ..., r_6$, in which r_i is given by:

$$r_1 = 1, r_2 = 1, r_{3=001},$$
$$r_4 = 01, r_5 = 0001, r_6 = 000001$$

As it can be seen, each "1" in c^{17} is equivalent to occurrence of a run in x^{18} (in fact, the number of runs in x^{18} is equal to the number of ones in c^{17} plus 1). The length of each run also equals the length of equivalent subsequence $r_1, r_2, ..., r_6$.

To examine the randomness of runs in an IRS, we first state the following theorem mentioned in [3].

Theorem 1. Suppose $X^n = X_1, X_2, ..., X_n$ is an IRS of length *n*. If C^{n-1} is constructed as follows:

$$\begin{aligned} C^{n-1} &= C_1, C_2, ..., C_{n-1} \\ &= X_1 \oplus X_2, X_2 \oplus X_3, ..., X_{n-1} \oplus X_n \end{aligned} \qquad (3)$$

Then C^{n-1} is also an IRS of length $n-1$.

According to the definition 2, we can consider each change from "1" to "0" or from "0" to "1" equivalent to occurrence of a run in the sequence which in turn is equivalent to occurrence of "1" in C^{n-1}. Therefore, it can readily be seen that each "1" in C^{n-1} corresponds to a run in x^{18}. So the random variable N_r, defined in (4), states the total number of runs in x^{18}.

$$N_r = \sum_{i=1}^{n-1} C_i + 1 \qquad (4)$$

It must be noticed that the number of runs in X^n is always equal to the number of ones in C^{n-1} plus 1 which is completely consistent with (4). Now we obtain the exact mean and variance of the total number of runs N_r.

Theorem 2. If $X^n = X_1, X_2, ..., X_n$ is an IRS of length n, then the mean and variance of its runs is $\frac{n+1}{2}$ and $\frac{n-1}{4}$ respectively.

Proof. We first construct C^{n-1} using (3). Since the components of C^{n-1} are independent and identically distributed (i.i.d), the mean and variance of random variables $C_i (i = 1,2,..., n-1)$ is:

$$E(C_i) = 1 \times P(C_i = 1) + 0 \times P(C_i = 0) = 1 \times \frac{1}{2} + 0 \times \frac{1}{2} = \frac{1}{2} \quad (5)$$

$$Var(C_i) = E(C_i - \frac{1}{2})^2 = (1-\frac{1}{2})^2 P(C_i = 1) + (0-\frac{1}{2})^2 P(C_i = 0) = \frac{1}{4} \quad (6)$$

Using (5) and (6), the mean and variance of N_r is simply calculated as:

$$E(N_r) = E(\sum_{i=1}^{n-1} C_i + 1) = (n-1) \times \frac{1}{2} + 1 = \frac{n+1}{2} \quad (7)$$

$$Var(N_r) = Var(\sum_{i=1}^{n-1} C_i + 1) = \sum_{i=1}^{n-1} Var(C_i) = \frac{n-1}{4} \quad (8)$$

Since the random variables $C_i (i = 1,2,..., n-1)$ are iid, according to Central Limit theorem, for sufficiently large amount of *n*, N_r tends to Normal distribution with $\frac{n+1}{2}$ and $\frac{n-1}{4}$ as its mean and variance, respectively. Thus the test statistic in (2) has Normal distribution for large enough *n*. The improved test statistic is then as follows:

$$T_n(obs) = \frac{N_{obs} - \frac{n+1}{2}}{\sqrt{\frac{n-1}{4}}} \quad (9)$$

In which, N_{obs} is the number of runs in the sequence under the test. According to *Hypothesis Testing* theory, the probability value (p-value or *PV*) is then found as:

$$PV = 2 \times (1 - \Phi_n(|T_n(obs)|)) \quad (10)$$

In (10), $\Phi_n(.)$ is the Standard Normal distribution function and *PV* is twice the area under the curve of Standard Normal probability density function from $|T_n(obs)|$ to infinity [6]. To perform the test it is sufficient to compute the *PV* using (10) and then to compare it with a level of significance α that is recommended to be $0.001 \le \alpha \le 0.01$ [6]. If the *PV* is equal to or greater than α then the sequence is considered random and if it is less than α then the sequence is reported as nonrandom. For the test to be accurate, *n* must be large enough. In practice, it is recommended to choose $n \ge 20$.

Probability Distribution Function of Length of Runs

To calculate the probability function for the lengths of runs in an IRS like $X^n = X_1, X_2, ..., X_n$, we first construct C^{n-1} using (3). As discussed in the previous section, each transition from "1" to "0" and vice versa corresponds to the occurrence of a run in X^n which is in turn equivalent to appearing of "1" in C^{n-1}. In fact the end of a run in X^n is specified by appearing "1" in C^{n-1} and the length of the run is equal to the distance of this "1" from the previous "1" in C^{n-1} (the length of firs run equals the position of firs "1" in C^{n-1}).

Now we define random variable T_k as the *time of kth victory*:

$$\begin{cases} T_0 = 0 \\ T_k = \min\{i \succ T_{k-1} : C_i = 1\} \\ k = 1,2,3,... \end{cases} \quad (11)$$

The above definition for T_k implies that it determines the position of the *k*th occurrence of "1" in C^{n-1} (the time of *k*th victory). Therefore random variables $R_1 = T_1 - T_0$, $R_2 = T2 - T1$, $R3 = T3 - T2$, … indicates the length of first run, the length of second run, the length of third run and so on, respectively.

Theorem 3. Random variables $R_1 = T_1 - T_0$, $R_2 = T2 - T1$, $R3 = T3 - T2$, …are independent and have identical Geometric distribution with parameter $p = \frac{1}{2}$.

Proof. To prove that random variables R_i are independent, we show that the joint probability mass function of them equals the product of probability mass function of each R_i. For $k \ge 1$ and $r_1, r_2, ..., r_k \ge 1$ we have:

$$\begin{aligned} &P(T_1 - T_0 = r_1, T_2 - T_1 = r_2, ...T_k - T_{k-1} = r_k) \\ &P(R_1 = r_1, R_2 = r_2, ..., R_k = r_k) = \\ &P(C_1 = 0, ..., C_{r_1 - 1} = 0, C_{r_1} = 1, \\ &C_{r_1+1} = 0, ..., C_{r_1+r_2-1} = 0, C_{r_1+r_2} = 1, ..., \\ &C_{\left(\sum_{j=1}^{k-1} r_j\right)+1} = 0, ..., C_{\left(\sum_{j=1}^{k} r_j\right)-1} = 0, C_{\sum_{j=1}^{k} r_j} = 1) \end{aligned} \quad (12)$$

We also know that random variables $C_i (i = 1,2,..., n-1)$ are i.i.d, so (12) will be written as:

$$P(R_1 = r_1, R_2 = r_2, ..., R_k = r_k)$$
$$= ((\frac{1}{2})^{r_1-1} \times \frac{1}{2}) \times ((\frac{1}{2})^{r_2-1} \times \frac{1}{2}) \times ... \times ((\frac{1}{2})^{r_k-1} \times \frac{1}{2}) \quad (13)$$
$$= P(R_1 = r_1)P(R_2 = r_2)...P(R_k = r_k)$$

Therefore the joint probability mass function of random variables $R_i (i = 1,2,...,k)$ equals the product of probability mass function of each one which is the Geometric probability distribution with parameter $p = \frac{1}{2}$, so each R_i is independent from others and we have:

$$P(R_1 = j) = P(R_2 = j) = ... = P(R_k = j)$$
$$= P(R = j) = (\frac{1}{2})^j \qquad j = 1,2,3,... \quad (14)$$

According to theorem 3, the probability of occurrence of a run of specific length is only determined by its length and is independent of its position or time of occurrence. In fact the lengths of runs are i.i.d with Geometric distribution with parameter $p = \frac{1}{2}$. From now on in this paper, we refer to R_i with symbol R , hence:

$$E(R) = \frac{1}{p} = 2 \quad (15)$$

$$Var(R) = \frac{1-p}{p} = 2 \quad (16)$$

In other words, the mean and the variance of the length of runs in an IRS are $\frac{1}{2}$.

Since in an IRS, the probability of "1" and "0" is the same and equal to $\frac{1}{2}$, a run of length j is a gap or a block with probability $\frac{1}{2}$. In other words, if $P(G = j)$ is the probability of occurrence of a j-bit gap and $P(B = j)$ is that of a j-bit block then using (4):

$$P(G = j) = P(B = j) = (\frac{1}{2})^{j+1} \qquad j = 1,2,3,... \quad (17)$$

And also:

$$P(R = j) = P(G = j) + P(B = j) = (\frac{1}{2})^j$$

It is worth mentioning that in [1] the above probability functions are calculated by means of information theoretic concepts.

Statistical Test on Length of Runs

In this section, two new statistical tests using the distribution of runs of different lengths, computed in Theorem 3, are presented. To perform these tests on an n-bit binary sequence, first the sequence R^k is constructed, using (11) and $R_i = T_i - T_{i-1}$:

$$X^n = X_1, X_2, ..., X_n$$
$$R^k = R_1, R_2, ..., R_k \quad (18)$$

We can now perform the test in two different types, *Test 1* and *Test 2*.

Test 1

According to theorem 2, the average of k is nearly $(n+1)/2$. First, suppose that we want to do the test only on runs of lengths of at most m. For this reason, it is assumed that the runs of length greater than m are equivalent to the occurrence of a hypothetic symbol *, whose probability is computed as follows. Using (14):

$$P(R_i = *) = P(R_i \succ m) = 1 - P(R_i \le m)$$
$$= 1 - \sum_{j=1}^{m} (\frac{1}{2})^j = 1 - (\frac{1}{2} + \frac{1}{4} + ... + \frac{1}{2^m}) = \frac{1}{2^m} \quad (19)$$

Thus we have:

$$P(R_i = j) = \begin{cases} (\frac{1}{2})^j & j = 1,2,...,m \\ \frac{1}{2^m} & j = * \end{cases} \quad (20)$$
$$i = 1,2,...,k$$

So it can be said that the sequence R^k corresponds a polynomial random variable with probability function given in (20). Using the theorem of chi-square test [9], the test statistic is as follows:

$$T_k(obs_r) = \sum_{i=1}^{m} \frac{(N_i - k \times 2^{-i})}{k \times 2^{-i}} + \frac{(N^* - k \times 2^{-m})^2}{k \times 2^{-m}} \quad (21)$$

In (21), N_i is the number of runs of length i and N^* is the total number of runs of lengths greater than m in the sequence X^n i.e. $N = \#\{R_i \mid R_i > m\}$. According to the theorem of chi-square test, $T_k(obs_r)$ in (21) has the chi-square distribution with m degrees of freedom. The PV is then calculated as:

$$PV = \text{igamc}(m/2, T_k(obs_r)/2) \quad (22)$$

In fact the above *PV* equals the area under the curve of probability density function of the chi-square random variable from $T_k(obs_r)$ to infinity. The igamc is the *Incomplete Gamma Function* defined in [6]. The final step to complete the test is to compare the *PV* to a level of significance α in order to decide if the sequence under the test is random or not, as discussed in section 3. For the test to be accurate, taking the approximations made in chi-square distribution into account [9], *n* must satisfy the following conditions:

$$\frac{n+1}{2}\times\frac{1}{2^m}\succ 5 \quad \text{or} \quad n\succ 5\times 2^{m+1}-1 \tag{23}$$

Or alternatively *m* must meet the following restraint:

$$m\prec \log_2(\frac{n+1}{10}) \tag{24}$$

Test 2

If one wants to perform the test on runs of lengths of at most *m* but does not want to consider the hypothetic symbol *, then the test can be done as follows. Using (18), the sequence $R^L = R_1, R_2, ..., R_L$ must first be constructed from X^n, in which R^L comprises only runs of lengths of at most *m*. Thus the probability function of random variables R_i $(i = 1,2,...,L)$ is conditional and computed as:

$$\begin{aligned} P(R_i = j \mid j \le m) &= \frac{P(R_i = j, j \le m)}{P(j \le m)} \\ &= \frac{P(R_i = j)}{\frac{1}{2}+\frac{1}{4}+...+\frac{1}{2^m}} = \frac{2^{-j}}{1-2^{-m}} \end{aligned} \tag{25}$$

It is clear in this case that:

$$\sum_{j=1}^{m} P(R_i = j \mid j \le m) = \sum_{j=1}^{m}\frac{2^{-i}}{1-2^{-m}} = 1 \tag{26}$$

In other words in the sequence R^L, every R_i has the probability function as follows:

$$P(R_i = j) = \frac{2^{-j}}{1-2^{-m}} \qquad j = 1,2,...,m \quad i = 1,...,L \tag{27}$$

Therefore the test statistic for runs test is:

$$T_L(obs_r) = \sum_{i=1}^{m}\frac{(r_i - L\times\frac{2^{-i}}{1-2^{-m}})^2}{L\times\frac{2^{-i}}{1-2^{-m}}} \tag{28}$$

The test statistic in (28) has the chi-square distribution with $m-1$ degrees of freedom, hence the *PV* is:

$$PV = \text{igamc}\left((m-1)/2, T_L(obs_r)/2\right) \tag{29}$$

The final step to complete the test is to compare the *PV* to a level of significance α in order to decide if the sequence under the test is random or not, as discussed in section 3. In ideal case, *L* is on average equal to:

$$\begin{aligned} \frac{n+1}{2}\times\sum_{j=1}^{m} P(R_i = j) &= \frac{n+1}{2}\times(\frac{1}{2}+\frac{1}{4}+...+\frac{1}{2^m}) \\ &= \frac{n+1}{2}\times(1-2^{-m}) \end{aligned} \tag{30}$$

The condition for the test to be accurate is the same as in (23). So *m* must also meet the restraint in (24).

$$\frac{n+1}{2}(1-2^{-m})\times\frac{2^{-m}}{(1-2^{-5})}\succ 5 \Rightarrow n\succ 5\times 2^{m+1}-1 \tag{31}$$

Statistical Test on length of Gaps and Blocks

Employing the procedure used in section 5, we can readily offer separate tests on blocks and gaps. Though these tests might be less practical compared to the test presented in previous section, they can be useful in specific cases. Assume that only gaps (or alternatively blocks) are examined in a sequence of length *n*:

$$\begin{aligned} X^n &= X_1, X_2, ..., X_n \\ G^k &= G_1, G_2, ..., G_k \end{aligned} \tag{32}$$

Since the number of gaps in an IRS is on average equal to $\frac{1}{2}$, considering theorem 2, *k* is on average equal to $(n+1)/4$. Let $P(G_i = j)$ be the probability that *i*th gap G_i in the sequence G^k is of length *j*. Since blocks of the sequence X^n is omitted in the sequence G^k and using (17) we have:

$$\begin{aligned} P(G_i = j) &= P(R_i = j \mid \text{run is a gap}) \\ &= \frac{P(R_i = j, \text{run is a gap})}{P(\text{run is agap})} \\ &= \frac{(\frac{1}{2})^{j+1}}{(\frac{1}{2})} = (\frac{1}{2})^j \qquad j = 1,2,... \quad i = 1,...,k \end{aligned} \tag{33}$$

Analogous relation can also be extracted for blocks:

$$\begin{aligned} B^{k'} &= B_1, B_2, ..., B_{k'} \\ P(B_i = j) &= \frac{1}{2^j} \qquad j = 1,2,... \quad i = 1,2,...,k' \end{aligned} \tag{34}$$

Obviously, the probability functions in (33) and (34) are consistent with that in (14). Employing the results in section 5, the test statistics for gaps test and blocks test are as below, respectively:

$$T_k(obs_g)=\sum_{i=1}^{m}\frac{(g_i-k\times 2^{-i})}{k\times 2^{-i}}+\frac{(g^*-k\times 2^{-m})^2}{k\times 2^{-m}}$$
$$T_{k'}(obs_b)=\sum_{i=1}^{m}\frac{(b_i-k'\times 2^{-i})}{k'\times 2^{-i}}+\frac{(b^*-k'\times 2^{-m})^2}{k'\times 2^{-m}} \quad (35)$$

In which g_i and b_i are the numbers of gaps and blocks of length i, respectively. Also $g*$ and $b*$ are the total numbers of gaps and blocks of lengths greater than m, respectively. It must be noticed that for the tests to be accurate, the following condition must be satisfied:

$$k\times\frac{1}{2^m}\succ 5\Rightarrow\frac{n+1}{4}\times\frac{1}{2^m}\succ 5\Rightarrow n\succ 5\times 2^{m+1}-1 \quad (36)$$

Or equivalently $m\prec\log_2(\frac{n+1}{10})$, that is the same as (24).

Now let's perform the test for gaps and blocks of lengths of at most m, but not consider the hypothetic symbol *, hence neglecting $g*$ and $b*$. By analogy to relations (25) to (31), parameters of the tests are determined as relations in (37) to (39):

$$G^L=G_1,G_2,...,G_L$$
$$B^{L'}=B_1,B_2,...,B_{L'} \quad (37)$$

$$\begin{cases}P(G_i=j)=\dfrac{2^{-j}}{1-2^{-m}}\\ i=1,2,...,L\end{cases}\qquad j=1,2,...,m \quad (38)$$
$$\begin{cases}P(B_i=j)=\dfrac{2^{-j}}{1-2^{-m}}\\ i=1,2,...,L'\end{cases}$$

$$T_L(obs_g)=\sum_{i=1}^{m}\frac{(g_i-L\times\frac{2^{-i}}{1-2^{-m}})^2}{L\times\frac{2^{-i}}{1-2^{-m}}}$$
$$T_{L'}(obs_b)=\sum_{i=1}^{m}\frac{(b_i-L'\times\frac{2^{-i}}{1-2^{-m}})^2}{L'\times\frac{2^{-i}}{1-2^{-m}}} \quad (39)$$

$T_L(obs_g)$ and $T_{L'}(obs_b)$ are chi-square random variables with $m-1$ degrees of freedom. The condition for the tests to be accurate is the same as in (36).

The foregoing discussions give some diverse and complete tests for evaluating statistical behavior of runs, gaps and blocks which in turn are significantly more accurate than the test presented in [6,10].

Simulation

Empirical results for sample data

In this section we present some empirical results which have been computed by performing Test 1 and Test 2 on some well known binary sample data. These sequences are binary representations of π (Pi number), e (Neper number), square root of 2 and square root of 3, used in [6]. In table (1) *PV*s of Test 1, computed on the mentioned binary sequences, are given for various amount of *m* and *n=100000*. It shows that by increasing the amount of *m* the *PV* decreases and hence the test becomes more accurate, since for a single test and under the same circumstances the smaller the *PV*, the stronger the test. Table (2) is those of Test 2.

Table 1. Results of Test 1 for some well known random sequences.

Test 1	π	e	$\sqrt{2}$	$\sqrt{3}$
m=3	0.9859	0.0767	0.0054	0.6477
m=5	0.6338	0.1983	2.2e-4	0.1071
m=8	0.0724	0.0610	7.4e-5	0.0392

Table 2. Results of Test 2 for some well known random sequences.

Test 2	π	e	$\sqrt{2}$	$\sqrt{3}$
m=3	0.9405	0.0357	0.0019	0.7447
m=5	0.7115	0.1354	2.2e-4	0.1265
m=8	0.0455	0.2205	0.0042	0.0234

In table (3) there is a comparison between the results of Runs Test, presented by NIST [6] and the equivalent results of Test 1, presented in this paper. As it can be seen, the results of the Test 1 are much better than those of Runs Test [6].

Table 3. Comparison between results of Test 1 and Runs Test presented in [6].

Tests	π	e	$\sqrt{2}$	$\sqrt{3}$
Test 1	0.6338	0.1983	2.2e-4	0.1071
Test 2	0.7115	0.1354	2.2e-4	0.1265
Runs Test [6]	0.4193	0.5619	0.3134	0.2611

These results are evidence for the accuracy of the presented test which is based on the evaluation of the probability distribution function itself rather than being based only on the mean and variance of the probability function. The results in table (3) are for n=1000000 and m=5.

Test results for reference generators

Table (4) shows the test results for two reference generators, namely Blum-Blum-Schub (BBS) and natural noise generator1 using special diodes.

Table 4. Test results for BBS generator and natural noise generator using diodes.

Tests	π (Pi number)	e (Neper number)	sqrt of 2	sqrt of 3
Runs Test (presented in [6]) (P-value)	0.4193	0.5619	0.3134	0.2611
Test 1 (P-value)	0.0831	0.0463	1.0669e-5	2.6315e-4

We have used the method mentioned in [6] to compute the uniformity of the test and the proportion of sequences passing a test with $\hat{p} = 1 - \alpha = 0.99$ and sample size 1000.

Conclusion

In this paper, the randomness of runs and their distribution and subsequently those of gaps and blocks, for an *Ideal Random Sequence* (IRS) were analyzed. Using the proper combination of an IRS and its shifted version, we found a precise method for calculating the exact mean and variance of the total number of runs in an IRS and consequently improved the test statistic for Runs Test in [5]. Since a test based on evaluation of the probability distribution function itself is more accurate than a test based only on the mean and variance of the probability distribution function, we presented two new chi-square tests based on the probability distribution function of length of runs whose test statistics are of m and m-1 degrees of freedom. In section 5, the distribution of length of gaps and blocks in an IRS were also found in the same way and accordingly some statistical tests were proposed. However, the tests presented in section 5 are less practical. It was shown that the tests presented in this paper are more tough to pass than those in [6,10] and can also examine the randomness of gaps and blocks of the sequence under the test.

[1] This generator has been designed in Cryptography Research Laboratory at Isfahan University of Technology (IUT) using Diode No. NC3021 manufactured by Noise/COM company.

References

[1] M. Dakhilalian, "Analyzing statistical properties of runs and its application in some tests", 11th Iranian Electrical Engineering Conference, Shiraz University, May 2003.

[2] B. Peizari, M. Dakhilalian, "Improvement of runs test and its application on sub blocks ", 2nd Iranian Society of Cryptography Conference, Sharif University, September 2003.

[3] M. Dakhilalian, Statistical Analyzing of Pseudorandom Sequences and Design of Chaotic Generators, PhD Thesis in Electrical Engineering, Isfahan University of Technology, October 1998.

[4] S.W. Golomb, Shift Register Sequence, Holden-Day, San Fransisco, 1982.

[5] H. Beker, and F. Piper, Cipher System: The Protection of Communication, Northwood Book, London, 1982.

[6] A.L. Rukhin, J. Soto, Nechvatal J., M. Smid, E. Barker, S. Leigh, M. Levenson, M. Vangel, D. Banks, A. Heckert, J. Dray, and S. Vo, "A statistical test suite for random and pseudorandom number generators for cryptographic applications", NIST Special Publication 800-22, 15 May 2001. See http://csrc.nist.gov/rng/.

[7] T. W. Anderson, and D. A. Darling, "A test of goodness-of-fit", Journal of the American Statistical Association, vol. 49, No. 268, pp. 765-769, 1954.

[8] H. Gustafson, et al, “A computer package for measuring strength of encryption algorithms”, Journal of Computers & Security, vo.13 No.8, 687-697, 1994. See http://www.isi.qut.edu.au/resources/cryptx.

[9] D. E. Knuth, The Art of Computer Programming, Volume 2: Seminumerical Algorithms, 3rd ed., Addison-Wesley, Reading, Mass, 1998.

[10] P. L’Ecuyer, and R. Simard, TestU01: A Software Library in ANSI C for Empirical Testing of Random Number Generators, Software user’s guide, 2001. See http://www.iro.umontreal.ca/ simardr/testu01/tu01.html.

[11] G. Marsaglia, "DIEHARD: a battery of tests of randomness", 1996.
See http://stat.fsu.edu/geo/diehard.html.

[12] G. Marsaglia, and W. W. Tsang, "Some difficult-to-pass tests of randomness", Journal of Statistical Software 7, 3, 1-9, 2002.
See http://www.jstatsoft.org/v07/i03/tuftests.pdf.

[13] S. Wegenkittl , Empirical Testing of Pseudorandom Number Generators, Master of Science Thesis, Salzbufg University, 1996.

[14] S. Wegenkittl, Generalized-divergence and Frequency Analysis in Markov Chains. Ph.D. thesis, University of Salzburg, 1998.
See http://random.mat.sbg.ac.at/team/.

[15] P. L’Ecuyer, "Testing random number generators", Proc. Winter Simulation Conf. IEEE Press, 305-313, 1992.

[16] G. Marsaglia, "A current view of random number generators", In Computer Science and Statistics, 16th Symposium on the Interface. Elsevier Science Publishers, North-Holland, Amsterdam, 3-10, 1985.

[17] U. Maurer, "A universal statistical test for random bit generators", Journal of Cryptology 5, 2, 89-105, 1992.

[18] A. L. Rukhin, "Testing randomness: A suite of statistical procedures", Theory of Probability and Its Applications 45, 1, 111-132, 2001.

Mohammad Dakhilalian received the B.Sc. and Ph.D. degrees in Electrical Engineering from Isfahan University of Technology (IUT) in 1989 and 1998 respectively and M.Sc. degree in Electrical Engineering from Tarbiat Modarres University in 1993. He was an Assistant Professor of Faculty of Information & Communication Technology, Ministry of ICT, Tehran, Iran in 1999-2001. He joined IUT in 2001 and is an Assistant Professor in Electrical and Computer Engineering Department. His current research interests are Cryptography and Data Security.

.

Ebrahim Molavian Jazi received BS degrees in Electrical Engineering and also Applied Mathematics in 2007 from Isfahan University of Technology (IUT), Isfahan, Iran, where he was a member of the Cryptography & Information Security research group. Currently, he is a graduate student in Electrical Engineering at the University of Notre Dame, IN, USA. His research interests include communications, information theory, and information security.

Ebrahim Molavian Jazi received his B.Sc. degree in Electrical Engineering from Isfahan University of Technology (IUT), Isfahan, Iran, in 2007. He was a member of Cryptography & Information Security Research Group at IUT from 2005 to 2007. He is currently a graduate student at Simon Fraser University (SFU), BC, Canada. His research interests include Communication Theory, Cryptography, Information Hiding, and Coding.