The first return time test of pseudorandom numbers

Dong Han Kim
Department of Mathematics
Korea Advanced Institute of Science and Technology
373-1 Kuseong-dong, Yuseong-ku
Taejon, 305-701 Korea
kim@euclid.kaist.ac.kr

1. Introduction

We introduce a new method of testing PRNGs based on the first return time of the some fixed length blocks in a randomly generated binary sequence. The first return time is closely related to entropy, which is the central idea in the information theory founded by C. Shannon (1948). For a binary source it is defined to be the limit of $-\frac{1}{n}\sum_{i=1}^{2^n} p_i \log p_i$ as n increases to infinity where the p_i's are the relative frequencies of 2^n blocks of length n in a typical binary sequence generated by the source. Entropy measures the information content or the amount of randomness. In data compression the entropy measures the maximum compression rate. If there are more patterns, that is, less randomness in a given sequence, then it has smaller entropy and can be compressed more.

In this article the first return time in a random binary sequence is investigated. Consider a stationary ergodic binary process on the space of infinite sequence $(\{0,1\}^\infty, \mu)$, where μ is the shift invariant ergodic probability measure on the σ-field generated by finite dimensional cylinders. For each sample sequence x define the first return time by

$$R_n(x) = \min\{j \geq 1 : x_1^n = x_{j+1}^{j+n}\}.$$

A.D. Wyner and Ziv (1989) proved that $\frac{1}{n}\log R_n(x)$ converges to the entropy of the sequence in measure and Ornstein and Weiss (1993) showed that the convergence is pointwise. Later, A.J. Wyner (1993) discovered that for a stationary aperiodic Markov chain with entropy h the random variable $\frac{1}{n}\log R_n$ has approximately a normal distribution with mean h in a suitable sense. For a sharp estimate of the convergence of the average of $\frac{1}{n}\log R_n$, see Choe and Kim (2000).

Kac's lemma (1947) implies that $E[R_n | x_1^n = a_1^n] = 1/\mu(a_1^n)$ for any finite string a_1^n with $\mu(a_1^n) > 0$. Since Kac's lemma implies $E[R_n] = \#\{a_1^n \in A^n : \mu(a_1^n) > 0\}$, we have

$$\lim_{n\to\infty} \frac{1}{n} \log E[R_n] = \text{topological entropy}.$$

This implies that the algorithm using $\log R_n$ is much more efficient than the algorithm of R_n.

Since Maurer's work (1992), the nonoverlapping first return time which corresponds to

$$R'_n(x) = \min\{j \geq 1 : x_1^n = x_{jn+1}^{jn+n}\},$$

has been investigated to be used in cryptography or in testing PRNGs. Nonoverlapping algorithm is relatively easier to analyze but overlapping method is more efficient and natural.

2. The probability distribution of the first return time

A block is a finite sequence of elements of A and an n-block is a block of length n. For an n-block $B = b_1 b_2 \cdots b_n$, we write $B_i^j = b_i b_{i+1} \cdots b_j, 1 \leq i \leq j \leq n$. Since the distribution of return time is different from block to block. We classify the blocks to each set of blocks have the same return time distribution.

Definition 1. Let B be an n-block. Suppose m satisfies $1 \leq m < n$ and

$$B_{m+1}^n = B_1^{n-m}.$$

The smallest such m is denoted by $\lambda_1(B)$, and the next smallest such m which is not a multiple of $\lambda_1(B)$ is called $\lambda_2(B)$, and we define $\lambda_k(B)$ by the smallest such m which is not a multiple of $\lambda_i(B)$ for every $i < k$. Let $\Lambda(B) = \{\lambda_1(B), \lambda_2(B), \dots\}$.

Definition 2. For an n-block B let $\mathcal{F}(B,k)$ and $\mathcal{S}(B,k)$ be the set of k-blocks defined by

$$\mathcal{F}(B,k) = \{C : C_1^n = B, C_{i+1}^{i+n} \neq B \text{ for any } i \geq 1\}, \quad k \geq n$$
$$\mathcal{S}(B,k) = \{C : (CB)_1^n = B, (CB)_{i+1}^{i+n} \neq B \text{ for any } i, 1 \leq i < k\}, \quad k \geq 1.$$

Proposition 1. *For all n-block B*

$$\mathcal{S}(B,k) = \mathcal{F}(B,k) \setminus \bigcup_{\lambda \in \Lambda(B)} \{C \in \mathcal{F}(B,k) : C_{k-\lambda+1}^k = B_1^\lambda\},$$

where the union is disjoint union. For $\lambda \in \Lambda(B)$ and $\ell = \max\{j \in \mathbb{N} : j\lambda < n\}$ we have

$$\{C \in \mathcal{F}(B,k) : C_{k-\lambda+1}^k = B_1^\lambda\} = \bigcup_{j=1}^{\ell} \{C \underbrace{B_1^\lambda \cdots B_1^\lambda}_{j} : C \in \mathcal{S}(B, k - j\lambda)\}.$$

Note that for any n-block B, we have $\lambda_i(B) > n/2$, $i \geq 2$ and $\ell = 1$ except $\lambda = \lambda_1(B)$.

Definition 3. Define $r_k(B)$ and $s_k(B)$ by

$$r_k(B) = \Pr(x_1^k \in \mathcal{F}(B,k)), \qquad s_k(B) = \Pr(x_1^k \in \mathcal{S}(B,k)).$$

From now on we consider i.i.d. processes. Since for every n-block B we have

$$\Pr(R_n(x) > k-n, x_1^n = B) = r_k(B), \qquad \Pr(R_n(x) = k, x_1^n = B) = s_k(B)\mu(B),$$

We can calculate the distribution of the first return time from the followings:

Proposition 2. *For i.i.d. processes, if $k > n$, we have*

$$r_k(B) = r_{k-1}(B) - \mu(B)s_{k-n}(B).$$

Let $m = |\Lambda(B)|$ and $\ell = \max\{i : i \cdot \lambda_1(B) < n\}$. For $k \geq n$

$$s_k(B) = r_k(B) - \sum_{i=1}^{\ell} \mu(B_1^{\lambda_1(B)})^i s_{k-i\cdot\lambda_1(B)}(B) - \sum_{i=2}^{m} \mu(B_1^{\lambda_i(B)}) s_{k-\lambda_i(B)}(B).$$

And we have $r_n(B) = \mu(B)$ and $s_k(B) = \begin{cases} 0 & \text{if } k \notin \Lambda(B), \\ \mu(B_1^k) & \text{if } k \in \Lambda(B), \end{cases}$ *for $k < n$.*

3. Application: test for the pseudorandom number generators

We calculate $\Pr(R_n = k)$ for every integer $k \geq 1$ numerically by using the formula in the previous section. The averages and the standard deviations of the logarithm of the first return time are computed and the deviation of the experimental data from the theoretical prediction is used to test PRNGs. We apply the standard Z-test for sample mean of $\frac{1}{n}\log R_n$ of each block. In each case the sample size is 100,000. The test result is shown in Table 1. Test A is to consider the Z-values for each 14-block. Since the Z-values among each blocks are highly correlated, we need to reduce the correlation among the Z-values. A binary block can be regarded as an integer in binary expansion so we say that a block is of type I (respectively II, III and IV), if the integer obtained by the binary block is 41 (mod 53) (respectively 59 (mod 89), 31 (mod 73) and 61 (mod 101)). Test B-I, II, III, IV is the variance test of the Z-values over the blocks of type I, II, III, IV. The experiments show that correlations among the blocks of type I, II, III and IV are negligible. The symbols $\triangle$ and $\times$ denote the cases when the corresponding generators fail the test with statistical confidence of 95% and 99%, respectively.

A linear congruential generator $LCG(M, a, b)$ is of the form $X_{n+1} \equiv aX_n + b \pmod{M}$. **Randu**, **ANSI**, **MS**, **Fishman** and **Ran0** are $LCG(2^{31}, 65539, 0)$, $LCG(2^{31}, 1103515245, 12345)$, $LCG(2^{31}, 214013, 2531011)$, $LCG(2^{31}-1, 950706376, 0)$ and $LCG(2^{31}-1, 16807, 0)$, respectively. **ICG** is an inversive congruential generator given by $X_{n+1} \equiv X_n^{-1} + 1 \pmod{2^{31}-1}$. **Ran1** is Ran0 with Bays-Durham shuffle. **Ran2** is L'Ecuyer's generator made up of $LCG(2147483563, 40014, 0)$ and $LCG(2147483399, 40692, 0)$ with Bays-Durham shuffle. **Ran3** is of the form $X_n \equiv X_{n-55} - X_{n-24} \pmod{2^{31}}$. **F90** is Ran0 combined with the Marsaglia shift register $X_{n+1} = X_n(I \oplus L^{13})(I \oplus R^{17})(I \oplus L^5)$, where $\oplus$ denotes the binary exclusive-or operation and L is the left-shift and R is the right-shift.

Table 1. The test result for $n = 14$

Generator	Test A	Test B-I	Test B-II	Test B-III	Test B-IV
Randu	×	-	-	-	-
ANSI	×	-	-	-	-
MS	×	-	-	-	-
Fishman		×	△	×	△
ICG					
Ran0		×	×	×	×
Ran1		×	×		×
Ran2					
Ran3					
F90					

REFERENCES

Choe, G.H. and Kim, D.H. (2000). Average convergence rate of the first return time. *Colloq. Math.*, **84/85**, 159-171.

Kac, M. (1947). On the notion of recurrence in discrete stochastic processes. *Bull. Amer. Math. Soc.*, **53**, 1002-1010.

Maurer, U. (1992). A universal statistical test for random bit generators. *J. Cryptology.*, **5**, 89-105.

Ornstein, D. and Weiss, B. (1993). Entropy and data compression schemes. *IEEE Trans. Inform. Theory.*, **39**, 78-83.

Shannon, C. (1948). The mathematical theory of communication. *Bell Sys. Tech. J.*, **27**, 379-423 and 623-656.

Wyner, A.D. and Ziv, J. (1989). Some asymptotic properties of the entropy of stationary ergodic data source with applications to data compression. *IEEE Trans. Inform. Theory.*, **35**, 1250-1258.

Wyner, A.J. (1993). *Strong Matching Theorems and Applications to Data Compression and Statistics. Ph.D Thesis, Stanford University.*

RESUME

An algorithm for obtaining the probability distribution of the first return time R_n for the initial n-block with overlapping is presented and used to test pseudorandom number generators. The averages and the standard deviations of $\log R_n$ are computed theoretically and the Z-test is applied.