A New Statistical Randomness Test: Saturation Point Test

Fatih Sulak

Atılım University Mathematics Department, İncek, Ankara, Turkey.
Tel: +90 312 5868283, e-mail: fatih.sulak@atilim.edu.tr

Abstract—In this work, we propose a new statistical randomness test, the Saturation Point Test, which can be applied to integer sequences as well as binary sequences and is designed to increase the number of tests for short sequences. The subject of Saturation Point Test is the index of integer, denoted by SP, where all possible integers occur in the given sequence. We evaluate the probability $Pr(SP = t)$ using Stirling numbers of the second kind and give a procedure to produce a p-value using this probability. Moreover, we state a pseudocode for the new test and evaluate the subinterval probabilities to apply χ^2 goodness of fit test.

Keywords—Saturation Point, Statistical Randomness Test, Stirling Numbers of the Second Kind

1. Introduction

In cryptography, random numbers play an important role, but generation of random numbers for cryptographic purposes is a difficult task. Random numbers are ideally generated using true random sources, called true random number generators (TRNGs), which use a nondeterministic source to produce random numbers. On the one hand, generation of random numbers using TRNGs is inefficient and, on the other hand, it is difficult to store and transfer large number of random bits. Therefore, deterministic algorithms, which are called pseudorandom number generators (PRNGs), are preferred to TRNGs. PRNGs take a truly random binary sequence (seed) of length k and produce a periodic "random looking" binary sequence of length $l >> k$ [1]. The characteristics of PRNGs are different from TRNGs. First, PRNGs are efficient compared to TRNGs, taking shorter time to produce numbers. They are also deterministic, meaning that a given sequence of numbers is reproducible. PRNGs are periodic while TRNGs have no period.

A statistical randomness test is developed to test a null hypothesis (H_0) which states the input sequence is random. The test takes a binary sequence as an input and "accepts" or "rejects" the hypothesis. Randomness tests are probabilistic and there are two types of errors. If the data is random and H_0 is rejected, type I error is occurred and if the data is nonrandom and H_0 is accepted, type II error is occurred. The probability of a type I error is called the level of significance of the test and usually denoted by α. A statistical test produces a real number between 0 and 1 which is called p-value. If p-value $> \alpha$ then H_0 is accepted, otherwise it is rejected. The level of significance varies depending

Manuscript received...

on applications, and for cryptographic applications it is usually set to 0,01.

The output sequences of PRNGs should be random looking, therefore statistical analysis of PRNGs are essential. This process is accomplished by producing a sample sequence using the PRNG, and evaluating it by a statistical test suite. There are many statistical test suites [2], [3], [7], [8], [6] in the literature that include a collection of statistical randomness tests. As well as PRNGs the outputs of cryptographic primitives such as block ciphers and hash functions should be also random looking so that when the outputs are analyzed, predicting the algorithm should not be possible. Therefore, the evaluation of the outputs of the algorithms by statistical randomness tests is of great importance.

The test suites in the literature are designed to evaluate the randomness of PRNGs and sequences. In order to test block ciphers and hash functions which produce short sequences, a new evaluation method is proposed [9]. In that work the authors chose 7 statistical randomness tests among the tests in the NIST test suite [2], described an alternative evaluation method for block ciphers and hash functions and applied the new method to various algorithms.

In this work, we propose the Saturation Point Test to increase the number of tests for short sequences. We give a procedure to produce a p-value using combinatorial identities and a pseudocode for the new test. Moreover, we evaluate the subinterval probabilities to apply the method described in [9].

2. Preliminaries

Randomness tests are usually designed to measure the randomness properties of binary sequences as the tests in the NIST test suite [2]. However, randomness tests can also be applied to the integer sequences, and there are several such tests in the literature like Coupon Collector Test, Maximum of t Test, Poker Test and the like [3]. In this work, we define a new randomness test, Saturation Point Test, which can be applied to integer sequences. In order to apply the test to binary sequences we convert the binary sequence into an integer sequence.

Let $\{t_1, t_2, \ldots, t_s\}$ be an integer sequence where $0 \leq t_i \leq 2^k - 1$ for $i = 1, 2, \ldots, s$ and let Coverage be the number of different integers among t_i's and denoted by Cov.

Stirling number of the second kind is the number of different ways to partition a set with n elements into k non-empty subsets and is denoted by $\left\{ n \atop k \right\}$ [4].

Example 1: Let us evaluate $\left\{ n \atop n-1 \right\}$. In order to find this number we need to find the number of different ways to partition a set with n elements into $n-1$ non-empty subsets, which means that we need to divide it into $n-2$ sets of size 1 and one set of size 2. For this purpose we need to choose two elements, and the rest is uniquely determined. Therefore, $\left\{ n \atop n-1 \right\} = \binom{n}{2}$.

Example 2: Let us evaluate $\left\{ n \atop 2 \right\}$. In order to find this number we need to find the number of different ways to partition a set with n elements into 2 non-empty subsets, which means that we first need to choose a subset of size s with $1 \leq s \leq n-1$. There are $2^n - 2$ different ways to accomplish this, but as we need unordered pairs we need to divide this number by 2. As a result $\left\{ n \atop 2 \right\} = 2^{n-1} - 1$.

***Theorem** 3:* [3] Let $\{t_1, t_2, \ldots, t_s\}$ be a sequence of integers where $0 \leq t_i \leq 2^k - 1$ for $i = 1, 2, \ldots, s$, and let Cov be the number of different integers among t_i's. Then

$$Pr(Cov = t) = \frac{2^k \cdot (2^k - 1) \cdots (2^k - (t-1))}{2^{ks}} \left\{ {s \atop t} \right\},$$

where $\left\{ {s \atop t} \right\}$ denote the Stirling number of the second kind.

Proof: Let a_j be the j^{th} different number appearing in the sequence $\{t_1, t_2, \ldots, t_s\}$, and let A_j be the set of indices that correspond to a_j. For example if the integer sequence is $\{4, 2, 2, 1, 4, 1, 2\}$ then $a_1 = 4$, $a_2 = 2$, $a_3 = 1$, $A_1 = \{1, 5\}$, $A_2 = \{2, 3, 7\}$, and $A_3 = \{4, 6\}$. Since $Cov = t$, we should consider t non-empty sets of $\{1, 2, \ldots, s\}$. The number of such arrangements is $\left\{ {s \atop t} \right\}$. Moreover, we can choose a_j in $2^k - j + 1$ different ways for $1 \leq j \leq t$. Therefore, the number of all possible arrangements is

$$2^k \cdot (2^k - 1) \cdots (2^k - (t-1)) \left\{ {s \atop t} \right\}.$$

In order to find $Pr(Cov = t)$ we should divide this number to all possible cases, that is $(2^k)^s$, and the result follows. □

3. Saturation Point Test

The subject of Saturation Point Test is the index of integer, denoted by SP, where all possible integers occur in the given sequence. We state the following theorem to determine the p-value and the subinterval probabilities for the test.

***Theorem** 4:* Let $\{a_1, a_2, \ldots, a_s\}$ be a sequence of integers where $0 \leq a_i \leq 2^k - 1$ for $i = 1, 2, \ldots, s$, and let SP be the index of integer where all possible integers occur in the sequence (if all integer do not occur then $SP = \infty$). Then

$$\begin{aligned} Pr(SP = t) &= \frac{2^k!}{2^{kt}} \left\{ {t-1 \atop 2^k - 1} \right\}, \\ Pr(SP = \infty) &= 1 - \sum_{i=2^k}^{s} Pr(SP = i) \end{aligned}$$

Proof: Assume that $SP = t$, then the coverage of the sequence for first $t-1$ integers should be $2^k - 1$ and the integer at index t should be the remaining integer. Let K denote the number of integers in the sequence, then

$$\begin{aligned} & Pr(SP = t) \\ &= Pr(Cov = 2^k - 1 | K = t - 1) \cdot \frac{1}{2^k} \\ &= \frac{2^k \cdot (2^k - 1) \cdots (2^k - (2^k - 1 - 1))}{2^{k(t-1)}} \\ & \quad \cdot \left\{ {t-1 \atop 2^k - 1} \right\} \cdot \frac{1}{2^k} \\ &= \frac{2^k!}{2^{kt}} \left\{ {t-1 \atop 2^k - 1} \right\} \end{aligned}$$

is obtained.

We should add all $Pr(SP = i)$ for $2^k \leq i \leq s$ and subtract this number from 1 to find $Pr(SP = \infty)$, thus we have

$$Pr(SP = \infty) = 1 - \sum_{i=2^k}^{s} Pr(SP = i).$$

□

In order to apply the test, an n-bit binary sequence $\{a_1, a_2, \ldots, a_n\}$ is divided into k-bit blocks and the corresponding integer values of the subsequences are evaluated (the remaining bits are discarded). Then, an integer sequence of $\{t_1, t_2, \ldots, t_{\lfloor n/k \rfloor}\}$ with $0 \leq t_i \leq 2^k - 1$ is tested for $1 \leq i \leq \lfloor n/k \rfloor$. SP of the sequence is determined and p-value for the test is obtained using Table 2. The pseudocode of the test is stated in Algorithm 3.1.

Algorithm 3.1: S. POINT TEST($\{a_n\}, k$)

```
for $i \leftarrow 1$ to $\lfloor \frac{n}{k} \rfloor$
  do
    {$t_i = \sum_{j=1}^{k} 2^{k-j} a_{j+(i-1)k}$;
for $i \leftarrow 1$ to $2^k$
  do
    {$index[i] = \lfloor \frac{n}{k} \rfloor + 1$;
     comment: $initialization\ of$
     $index\ array$
$SP = 1$;
for $i \leftarrow 1$ to $2^k$
  do
    {for $j \leftarrow 1$ to $\lfloor \frac{n}{k} \rfloor$
       do
         {if $t_j = i$
            then $index[i] = j$; break
for $i \leftarrow 1$ to $2^k$
  do
    {if $index[i] > SP$
       then $SP = index[i]$;
$Use\ table\ 2\ to\ determine$
$the\ p - value$;
return $(p - value)$
```

In a set of t elements, the expected value of Cov is $\Theta(t \log(t))$ [5]. Therefore, we should have

$$\left\lfloor \frac{n}{k} \right\rfloor \geq 2^k \cdot k.$$

For $n = 256$ we suggest to choose $k = 4$. The expected value of SP is 51, thus we assign p-value=1 for $SP = 51$ and determine the other p-values according to their probabilities using Theorem 4. We also calculate the subinterval probabilities using Theorem 4, and state the results in Table 1. In that case the sequence consist of 64 integers. All integers may occur soonest in 16^{th} index, therefore Obs Min value can be minimum 16. It is also possible that all the integers may not occur in the sequence, in such a case we represent it by ∞ in the table.

TABLE 1
Subinterval Probabilities for Saturation Point Test

Sat Point		
Obs Min	Obs Max	Prob
16	38	0,193609
39	45	0,179686
46	53	0,196007
54	64	0,195881
∞		0,234818

Table 1 can be used to test cryptographic primitives like block ciphers and hash functions which produce 256-bit outputs. For this purpose, first an output collection of the cryptographic primitive which is the subject of the test is obtained, then this set is tested using the χ^2 goodness of fit method.

Following the same notation with [9], let m denote the number of output sequences of the cryptographic primitive, F_i denote the number of p-values in subinterval i, and p_i denote the probability of a p-value to be in subinterval i for $i = 1, 2, \ldots, 5$ in Table 1, then

$$\begin{aligned} \chi^2 &= \sum_{i=1}^{5} \frac{(F_i - m \cdot p_i)^2}{m \cdot p_i}, \\ p\text{-value} &= \texttt{igamc}\left(2, \frac{\chi^2}{2}\right), \end{aligned}$$

where `igamc` is the incomplete gamma function. If p-value$\geq$0.01, the cryptographic primitive which is the subject of the test is considered to be indistinguishable from a random mapping.

Poker Test and Coupon Collector Test also take Cov as subject [3]. Poker Test considers groups of five integers as in a poker game and observes the possible patterns; Coupon Collector Test observes the length of the sequence to have a complete set. Saturation Point Test is similar to these two tests as Cov is the subject, but unlike Poker Test and Coupon Collector Test, a p-value is produced for a sequence of 256-bit using Saturation Point Test.

4. Conclusion

In this work, we propose a new statistical test, Saturation Point Test, which can be applied to short binary sequences and integer sequences. We evaluate the probability $Pr(SP = t)$ using Stirling numbers of the second kind and give a procedure to produce a p-value according to this probability. We also state a pseudocode for the new test. Moreover, we give the subinterval probabilities to apply the evaluation method in [9] using the χ^2 goodness of fit method. Therefore, Saturation Point Test can be used to measure the randomness properties of cryptographic primitives like block ciphers and hash functions using the method described in [9].

As a future work, correlations between the other statistical randomness tests which can be applied to the short sequences, and Saturation Point Test can be analyzed.

References

[1] A. J. Menezes, P. C. van Oorschot, S. A. Vanstone, *Handbook of Applied Cryptography*, CRC Press, 2001.

[2] A. Rukhin, J. Soto, J. Nechvatal, M. Smid, E. Barker, S. Leigh, M. Levenson, M. Vangel, D. Banks, A. Heckert, J. Dray, S. Vo, *A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications*, NIST Special Publication 800-22, 2001

[3] D. E. Knuth, *Seminumerical Algorithms*, *The Art of Computer Programming*, vol 2, Addison-Wesley, 1981.

[4] R. L. Graham, D. E. Knuth, O. Patashnik, *Concrete Mathematics*, Addison-Wesley, 1988.

[5] G. Blom, L. Holst, D. Sandell, *Problems and Snapshots from the World of Probability*, Springer-Verlag, 1994.

[6] P. L'Ecuyer, R. Simard, *TestU01: A C library for empirical testing of random number generators*, ACM Trans. Math. Softw., vol. 33, no. 4, p.22, 2007.

[7] W. Caelli, E. Dawson, L. Nielsen, H. Gustafson, *CRYPT–X Statistical Package Manual, Measuring the strength of Stream and Block Ciphers*, Queensland University of Technology, 1992.

[8] G. Marsaglia, *The Marsaglia Random Number CDROM including the DIEHARD Battery of Tests of Randomness*, preprint, 1996. http://stat.fsu.edu/pub/diehard

[9] F. Sulak, A. Doğanaksoy, B. Ege, O. Koçak, *Evaluation of Randomness Test Results for Short Sequences*, Claude Carlet and Alexander Pott Ed., in Proc. Sixth Conference on Sequences and Their Applications. SETA 2010, Paris, 2010, vol. LNCS 6338, pp.309-319.

Appendix

TABLE 2
P-Value Table for the Saturation Point Test

T-value	p-value	T-value	p-value
16	0,000002	41	0,536917
17	0,000019	42	0,588901
18	0,000089	43	0,641373
19	0,000292	44	0,694029
20	0,000774	45	0,746589
21	0,001752	46	0,798799
22	0,003522	47	0,850431
23	0,006442	48	0,901284
24	0,010921	49	0,951183
25	0,017385	50	0,999979
26	0,026255	51	1,000000
27	0,037918	52	0,952454
28	0,052704	53	0,906218
29	0,070868	54	0,861397
30	0,092580	55	0,818054
31	0,117922	56	0,776233
32	0,146887	57	0,735966
33	0,179384	58	0,697270
34	0,215250	59	0,660150
35	0,254260	60	0,624599
36	0,296138	61	0,590604
37	0,340570	62	0,558141
38	0,387218	63	0,527182
39	0,435729	64	0,497693
40	0,485746	∞	0,469636