An Alternative Approach to Maurer's Universal Statistical Test

Ali DOĞANAKSOY, Cihangir TEZCAN

December 19, 2008

Table of contents

Randomness in Cryptography

Some uses of random numbers in Cryptography:

- **Keystreams of one-time pads**

Randomness in Cryptography

Some uses of random numbers in Cryptography:

- **Keystreams of one-time pads**
- **Secret keys of a symmetric cipher systems**

Randomness in Cryptography

Some uses of random numbers in Cryptography:

- **Keystreams of one-time pads**
- **Secret keys of a symmetric cipher systems**
- **Public key parameters**

Randomness in Cryptography

Some uses of random numbers in Cryptography:

- **Keystreams of one-time pads**
- **Secret keys of a symmetric cipher systems**
- **Public key parameters**
- **Session keys**

Randomness in Cryptography

Some uses of random numbers in Cryptography:

- **Keystreams of one-time pads**
- **Secret keys of a symmetric cipher systems**
- **Public key parameters**
- **Session keys**
- **Nonces**

Randomness in Cryptography

Some uses of random numbers in Cryptography:

- **Keystreams of one-time pads**
- **Secret keys of a symmetric cipher systems**
- **Public key parameters**
- **Session keys**
- **Nonces**
- **Initialization vectors**

Randomness in Cryptography

Some uses of random numbers in Cryptography:

- **Keystreams of one-time pads**
- **Secret keys of a symmetric cipher systems**
- **Public key parameters**
- **Session keys**
- **Nonces**
- **Initialization vectors**
- **Salts**

Frequency Test

- Calculates weight of the binary sequence.

Frequency Test

- Calculates weight of the binary sequence.
- Probability of weight being w is:

$$P(w) = \frac{1}{2^n}\binom{n}{w} \tag{1}$$

Frequency Test

- Calculates weight of the binary sequence.
- Probability of weight being w is:

$$P(w) = \frac{1}{2^n}\binom{n}{w} \tag{1}$$

$$p - value = \frac{1}{2^{n-1}}\sum_{i=1}^{w}\binom{n}{w} \tag{2}$$

Frequency Test

- Calculates weight of the binary sequence.
- Probability of weight being w is:

$$P(w) = \frac{1}{2^n}\binom{n}{w} \tag{1}$$

$$p - value = \frac{1}{2^{n-1}}\sum_{i=1}^{w}\binom{n}{w} \tag{2}$$

- Sequence:
  11111111111111111111000000000000000000000

Test Suites

- **NIST**
- **Knuth**
- **DIEHARD**
- **TestU01**
- **Crypt-X**

Maurer's Universal Statistical Test

- Introduced by Ueli M. Maurer in 1992.

Maurer's Universal Statistical Test

- Introduced by Ueli M. Maurer in 1992.
- Closely related to the per-bit entropy of a stream.

Maurer's Universal Statistical Test

- Introduced by Ueli M. Maurer in 1992.
- Closely related to the per-bit entropy of a stream.
- Designed to detect statistical defects that can be modeled by an ergodic design with finite memory.

Maurer's Universal Statistical Test

Sequence: 01011010011101010111

Maurer's Universal Statistical Test

Sequence: 01011010011101010111
L: 2

Maurer's Universal Statistical Test

Sequence: 01011010011101010111
L: 2
Sequence: 01 01 10 10 01 11 01 01 01 11

Maurer's Universal Statistical Test

Sequence: 01011010011101010111
L: 2
Sequence: 01 01 10 10 01 11 01 01 01 11
New Sequence: 1 1 2 2 1 3 1 1 1 3

Maurer's Universal Statistical Test

Sequence: 01011010011101010111
L: 2
Sequence: 01 01 10 10 01 11 01 01 01 11
New Sequence: 1 1 2 2 1 3 1 1 1 3
Distances: 3 6 2 1 1 4

Maurer's Universal Statistical Test

Sequence: 01011010011101010111
L: 2
Sequence: 01 01 10 10 01 11 01 01 01 11
New Sequence: 1 1 2 2 1 3 1 1 1 3
Distances: 3 6 2 1 1 4

$$f_n = \frac{1}{K}\sum_{1}^{K} \log_2 c_i = 1.1949875 \qquad (3)$$

Maurer's Universal Statistical Test

Sequence: 01011010011101010111
L: 2
Sequence: 01 01 10 10 01 11 01 01 01 11
New Sequence: 1 1 2 2 1 3 1 1 1 3
Distances: 3 6 2 1 1 4

$$f_n = \frac{1}{K}\sum_{1}^{K} \log_2 c_i = 1.1949875 \tag{3}$$

$$p - value = 0.767189 \tag{4}$$

Maurer's Universal Statistical Test

- Let $\{a_n\} = a_1, a_2, \ldots, a_N$ be a binary sequence of length N.

Maurer's Universal Statistical Test

- Let $\{a_n\} = a_1, a_2, \ldots, a_N$ be a binary sequence of length N.
- Partition this sequence into adjacent non-overlapping blocks of length L.

Maurer's Universal Statistical Test

- Let $\{a_n\} = a_1, a_2, \ldots, a_N$ be a binary sequence of length N.
- Partition this sequence into adjacent non-overlapping blocks of length L.
- Compute the integer values of these blocks and obtain a new sequence $\{t_n\} = t_1, t_2, \ldots, t_k$ where $k = \lfloor \frac{N}{L} \rfloor$ and $t_i \in \{0, 1, 2, \ldots, 2^L - 1\}$.

Maurer's Universal Statistical Test

- Let $\{a_n\} = a_1, a_2, \ldots, a_N$ be a binary sequence of length N.
- Partition this sequence into adjacent non-overlapping blocks of length L.
- Compute the integer values of these blocks and obtain a new sequence $\{t_n\} = t_1, t_2, \ldots, t_k$ where $k = \lfloor \frac{N}{L} \rfloor$ and $t_i \in \{0, 1, 2, \ldots, 2^L - 1\}$.
- The remaining bits at the end of the sequence are discarded.

Maurer's Universal Statistical Test

- The first Q blocks of $\{t_n\}$ is called the initialization part ($Q = 10 \cdot 2^L$).

Maurer's Universal Statistical Test

- The first Q blocks of $\{t_n\}$ is called the initialization part ($Q = 10 \cdot 2^L$).
- The remaining K blocks are called the test part where $K + Q = \lfloor n/L \rfloor$ ($K \approx 1000 \cdot 2^L$).

Maurer's Universal Statistical Test

- The first Q blocks of $\{t_n\}$ is called the initialization part ($Q = 10 \cdot 2^L$).
- The remaining K blocks are called the test part where $K + Q = \lfloor n/L \rfloor$ ($K \approx 1000 \cdot 2^L$).
- For every block in the test part, we calculate the distance of that block to its previous occurrence.

Maurer's Universal Statistical Test

- The first Q blocks of $\{t_n\}$ is called the initialization part ($Q = 10 \cdot 2^L$).
- The remaining K blocks are called the test part where $K + Q = \lfloor n/L \rfloor$ ($K \approx 1000 \cdot 2^L$).
- For every block in the test part, we calculate the distance of that block to its previous occurrence.
- If we denote these distances by c_i, the test statistic f_n is

Maurer's Universal Statistical Test

- The first Q blocks of $\{t_n\}$ is called the initialization part ($Q = 10 \cdot 2^L$).
- The remaining K blocks are called the test part where $K + Q = \lfloor n/L \rfloor$ ($K \approx 1000 \cdot 2^L$).
- For every block in the test part, we calculate the distance of that block to its previous occurrence.
- If we denote these distances by c_i, the test statistic f_n is

$$f_n = \frac{1}{K} \sum_{1}^{K} \log_2 c_i. \tag{5}$$

Maurer's Universal Statistical Test

The reference distribution for the test statistic is the half-normal distribution. The p-value is obtained as follows:

Maurer's Universal Statistical Test

The reference distribution for the test statistic is the half-normal distribution. The p-value is obtained as follows:

$$c = 0.7 - \frac{0.8}{L} + \left(4 + \frac{32}{L}\right) \frac{K^{-3/L}}{15} \tag{6}$$

Maurer's Universal Statistical Test

The reference distribution for the test statistic is the half-normal distribution. The p-value is obtained as follows:

$$c = 0.7 - \frac{0.8}{L} + \left(4 + \frac{32}{L}\right) \frac{K^{-3/L}}{15} \tag{6}$$

$$\sigma = c\sqrt{\frac{variance(L)}{K}} \tag{7}$$

Maurer's Universal Statistical Test

The reference distribution for the test statistic is the half-normal distribution. The p-value is obtained as follows:

$$c = 0.7 - \frac{0.8}{L} + \left(4 + \frac{32}{L}\right) \frac{K^{-3/L}}{15} \tag{6}$$

$$\sigma = c\sqrt{\frac{variance(L)}{K}} \tag{7}$$

$$p - value = erfc\left(\left|\frac{f_n - expectedvalue(L)}{\sqrt{2}\sigma}\right|\right) \tag{8}$$

Maurer's Universal Statistical Test

The reference distribution for the test statistic is the half-normal distribution. The p-value is obtained as follows:

$$c = 0.7 - \frac{0.8}{L} + \left(4 + \frac{32}{L}\right) \frac{K^{-3/L}}{15} \tag{6}$$

$$\sigma = c\sqrt{\frac{variance(L)}{K}} \tag{7}$$

$$p - value = erfc\left(\left|\frac{f_n - expectedvalue(L)}{\sqrt{2}\sigma}\right|\right) \tag{8}$$

If the obtained p-value is less than the probability of type I error, which is a small value between 0.01 and 0.001, we assume that the sequence is obtained from a non-random resource.

Parameters

Table: Parameters in NIST Statistical Test Suite

n	L	$Q = 10 \cdot 2^L$
$\geq 387,840$	6	640
$\geq 904,960$	7	1,280
$\geq 2,068,480$	8	2,560
$\geq 4,654,080$	9	5,120
$\geq 10,342,400$	10	10,240
$\geq 22,753,280$	11	20,480
$\geq 49,643,520$	12	40,960
$\geq 107,560,960$	13	81,920
$\geq 231,669,760$	14	163,840
$\geq 496,435,200$	15	327,680
$\geq 1,059,061,760$	16	655,360

Maurer's Universal Statistical Test

Sequence: 01011010011101010111

Maurer's Universal Statistical Test

Sequence: 01011010011101010111
L: 2

Maurer's Universal Statistical Test

Sequence: 01011010011101010111
L: 2
Sequence: 01 01 10 10 01 11 01 01 01 11

Maurer's Universal Statistical Test

Sequence: 01011010011101010111
L: 2
Sequence: 01 01 10 10 01 11 01 01 01 11
New Sequence: 1 1 2 2 1 3 1 1 1 3

Maurer's Universal Statistical Test

Sequence: 01011010011101010111
L: 2
Sequence: 01 01 10 10 01 11 01 01 01 11
New Sequence: 1 1 2 2 1 3 1 1 1 3
Distances: 0 0 0 1 3 6 2 1 1 4

An Alternative Approach

- Let $\{a_n\} = a_1, a_2, \ldots, a_N$ be a binary sequence of length N.

An Alternative Approach

- Let $\{a_n\} = a_1, a_2, \ldots, a_N$ be a binary sequence of length N.
- Partition $\{a_n\}$ into adjacent non-overlapping blocks of length L.

An Alternative Approach

- Let $\{a_n\} = a_1, a_2, \ldots, a_N$ be a binary sequence of length N.
- Partition $\{a_n\}$ into adjacent non-overlapping blocks of length L.
- Compute the integer values of these blocks.

An Alternative Approach

- Let $\{a_n\} = a_1, a_2, \ldots, a_N$ be a binary sequence of length N.
- Partition $\{a_n\}$ into adjacent non-overlapping blocks of length L.
- Compute the integer values of these blocks.
- We obtain a new sequence $\{t_n\} = t_1, t_2, \ldots, t_k$ where $k = \lfloor \frac{N}{L} \rfloor$ and $t_i \in \{0, 1, 2, \ldots, 2^L - 1\}$.

An Alternative Approach

- Let $\{a_n\} = a_1, a_2, \ldots, a_N$ be a binary sequence of length N.
- Partition $\{a_n\}$ into adjacent non-overlapping blocks of length L.
- Compute the integer values of these blocks.
- We obtain a new sequence $\{t_n\} = t_1, t_2, \ldots, t_k$ where $k = \lfloor \frac{N}{L} \rfloor$ and $t_i \in \{0, 1, 2, \ldots, 2^L - 1\}$.
- Obtain another sequence $\{c_n\}$ where c_i is the distance between the integer t_i and its previous occurrence.

An Alternative Approach

- Let $\{a_n\} = a_1, a_2, \ldots, a_N$ be a binary sequence of length N.
- Partition $\{a_n\}$ into adjacent non-overlapping blocks of length L.
- Compute the integer values of these blocks.
- We obtain a new sequence $\{t_n\} = t_1, t_2, \ldots, t_k$ where $k = \lfloor \frac{N}{L} \rfloor$ and $t_i \in \{0, 1, 2, \ldots, 2^L - 1\}$.
- Obtain another sequence $\{c_n\}$ where c_i is the distance between the integer t_i and its previous occurrence.
- If the integer t_i is its first occurrence in the sequence $\{t_n\}$, we assign the value 0 to c_i.

An Alternative Approach

If c_i is r for some i, this means that t_i and t_{i-r} are the same integers and the integers between them are different than t_i.

An Alternative Approach

If c_i is r for some i, this means that t_i and t_{i-r} are the same integers and the integers between them are different than t_i. The probability of such a situation is

$$prob(c_i = r) = \frac{(2^L - 1)^{r-1}}{(2^L)^r} \tag{9}$$

An Alternative Approach

If c_i is r for some i, this means that t_i and t_{i-r} are the same integers and the integers between them are different than t_i. The probability of such a situation is

$$prob(c_i = r) = \frac{(2^L - 1)^{r-1}}{(2^L)^r} \tag{9}$$

Note that since t_1 is the first element of the sequence $\{t_n\}$, c_1 must be 0. Similarly c_2 is either 0 or 1. Thus a distance r in the sequence c_i can be observed only in $k - r$ different places.

An Alternative Approach

If c_i is r for some i, this means that t_i and t_{i-r} are the same integers and the integers between them are different than t_i. The probability of such a situation is

$$prob(c_i = r) = \frac{(2^L - 1)^{r-1}}{(2^L)^r} \tag{9}$$

Note that since t_1 is the first element of the sequence $\{t_n\}$, c_1 must be 0. Similarly c_2 is either 0 or 1. Thus a distance r in the sequence c_i can be observed only in $k - r$ different places. Hence the expected number of appearance of the value r in the sequence $\{c_n\}$ is

$$E(r) = (k - r)\frac{(2^L - 1)^{r-1}}{(2^L)^r}. \tag{10}$$

An Alternative Approach

- The number of appearance of 0 in the $\{c_n\}$ sequence is equivalent to the number of distinct integers in the $\{t_n\}$ sequence.

An Alternative Approach

- The number of appearance of 0 in the $\{c_n\}$ sequence is equivalent to the number of distinct integers in the $\{t_n\}$ sequence.
- We will consider the cases when the probability of every possible integer values not appearing in $\{t_n\}$ is less than 10^{-4}.

An Alternative Approach

- The number of appearance of 0 in the $\{c_n\}$ sequence is equivalent to the number of distinct integers in the $\{t_n\}$ sequence.
- We will consider the cases when the probability of every possible integer values not appearing in $\{t_n\}$ is less than 10^{-4}.
- Hence to perform the test, the largest block size L that can be chosen for a $\{t_n\}$ sequence with length k is the largest L value satisfying the following equation:

An Alternative Approach

- The number of appearance of 0 in the $\{c_n\}$ sequence is equivalent to the number of distinct integers in the $\{t_n\}$ sequence.
- We will consider the cases when the probability of every possible integer values not appearing in $\{t_n\}$ is less than 10^{-4}.
- Hence to perform the test, the largest block size L that can be chosen for a $\{t_n\}$ sequence with length k is the largest L value satisfying the following equation:

$$1 - \left(\frac{2^L - 1}{2^L}\right)^k - \left(\frac{2^L - 2}{2^L}\right)^k - \ldots - \left(\frac{1}{2^L}\right)^k > 0.9999 \quad (11)$$

An Alternative Approach

Let d_i denote the number of appearance of the value i in the sequence $\{c_n\}$. We apply the test by calculating d_i's and performing χ^2 of goodness of fit test to d_i and $E(i)$ values. The degree of freedom d is $k-1$ and χ^2 value is

An Alternative Approach

Let d_i denote the number of appearance of the value i in the sequence $\{c_n\}$. We apply the test by calculating d_i's and performing χ^2 of goodness of fit test to d_i and $E(i)$ values. The degree of freedom d is $k-1$ and χ^2 value is

$$\chi^2 = \sum_{i=1}^{k} \frac{(d_i - E(i))^2}{E(i)} \tag{12}$$

An Alternative Approach

Let d_i denote the number of appearance of the value i in the sequence $\{c_n\}$. We apply the test by calculating d_i's and performing χ^2 of goodness of fit test to d_i and $E(i)$ values. The degree of freedom d is $k-1$ and χ^2 value is

$$\chi^2 = \sum_{i=1}^{k} \frac{(d_i - E(i))^2}{E(i)} \tag{12}$$

The p-value is:

$$p - value = gammq\left(\frac{d}{2}, \frac{\chi^2}{2}\right) \tag{13}$$

An Alternative Approach

Table: Largest Possible Block Sizes

Sequence Length	Block Size
$66 \leq n \leq 206$	2
$207 \leq n \leq 571$	3
$572 \leq n \leq 1,454$	4
$1,455 \leq n \leq 3,509$	5
$3,510 \leq n \leq 8,224$	6
$8,225 \leq n \leq 18,839$	7
$18,832 \leq n \leq 42,407$	8
$42,408 \leq n \leq 94,269$	9
$94,270 \leq n \leq 207,448$	10
$207,449 \leq n \leq 452,663$	11
$452,664 \leq n \leq 980,823$	12
$980,824 \leq n \leq 2,112,599$	13
$2,112,600 \leq n \leq 4,257,059$	14
$4,257,060 \leq n \leq 9,657,775$	15
$9,657,776 \leq n \leq 20,522,858$	16
$20,522,859 \leq n \leq 43,460,243$	17
$43,460,244 \leq n \leq 91,749,460$	18
$91,749,461 \leq n \leq 193,156,859$	19
$193,156,860 \leq n \leq 405,629,468$	20
$405,629,469 \leq n \leq 849,890,425$	21
$849,890,426 \leq n \leq 1,777,043,709$	22

Comparison

1. There is no initialization part in our method which allows us to test the whole sequence without wasting any parts of the sequence.

Comparison

1. There is no initialization part in our method which allows us to test the whole sequence without wasting any parts of the sequence.
2. In Maurer's Universal Test, sequences which are shorter than $387,840$ bits cannot be tested. However, our approach can be applied to test sequences as short as 66 bits.

Comparison

1. There is no initialization part in our method which allows us to test the whole sequence without wasting any parts of the sequence.
2. In Maurer's Universal Test, sequences which are shorter than $387,840$ bits cannot be tested. However, our approach can be applied to test sequences as short as 66 bits.
3. Maurer's test is not suitable for block size length larger than 16. This number is increased to 22 in our method.

Comparison

1. There is no initialization part in our method which allows us to test the whole sequence without wasting any parts of the sequence.
2. In Maurer's Universal Test, sequences which are shorter than $387,840$ bits cannot be tested. However, our approach can be applied to test sequences as short as 66 bits.
3. Maurer's test is not suitable for block size length larger than 16. This number is increased to 22 in our method.
4. Speed of the both algorithms:

Comparison

1. There is no initialization part in our method which allows us to test the whole sequence without wasting any parts of the sequence.
2. In Maurer's Universal Test, sequences which are shorter than $387,840$ bits cannot be tested. However, our approach can be applied to test sequences as short as 66 bits.
3. Maurer's test is not suitable for block size length larger than 16. This number is increased to 22 in our method.
4. Speed of the both algorithms:

Table: The time comparison of the two tests based on 1,000 random sequences of length 800,000 bits

Maurer's Universal Test	102 seconds
New method	87 seconds

Future Work

This new approach allows us to test short sequences and we are using it to test candidate Hash functions of NIST's Hash Function Competition.

Thank You

THANK YOU FOR YOUR ATTENTION