Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 22 (2013) 1201 – 1209

Procedia
Computer Science

17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems - KES2013

Randomness Criteria of the RMT-test Compared to the NIST

Yuuta Mikamori[a], Xin Yang[a] , Ryota Itoi, Mieko Tanaka-Yamawaki[a, *]

[a]Information and electronics Department, Graduate School of Engineering, Tottori University, Tottori 680-0945, Japan

Abstract

In this article, we report a new result of the error limit to be used for the RMT-test, which we have proposed earlier in order to measure the randomness of one-dimensional data sequence based on the comparison to the theoretical value derived by the random matrix theory (RMT). This new limit is obtained by comparing the error level of the RMT-test to the result of the NIST test. We prepared data sequences of various levels of randomness by shuffling a regular sequence many times. The result shows that the RMT error must be less than 0.60% in order to satisfy the requirement of the NIST test. This new limit is severer than the limit that we have obtained in the study of pseudo-random sequences. Although we need to consider the fact that the previous limit was the result of averaging over many samples, and the NIST test is applied only binary sequences and the conditions to apply the two tests are not the same, this result suggests us to reconsider the error limit of the RMT-test in more detail.

Keywords: Randomness, RMT-test, Evaluation criteria of randomness, NIST randomness test;

1. Introduction

The randomness means a degree of the difficulty at the time of predicting the next element in a sequence of numbers. In practice, however, a good tool to measure the randomness of a given data sequence is hard to find, just like the difficulty finding good random generators as well [1]. Many standard tools, such as JIS, NIST tools which are used in the field of cryptography [2], and so on, are bound to use a combination of multiple criteria to determine the randomness. Furthermore, a data type (binary, integer, or real numbers) and length are strongly restricted.

In order to overcome such difficulty, the authors have developed a new methodology, the RMT-test, by applying the random matrix theory (RMT) [3-5]. The advantage of the RMT-test is that it can measure the

* Corresponding author. Tel.: +81-0857-31-5223; fax: +81-0857-31-0879.
E-mail address: mieko@ike.tottori-u.ac.jp.

1877-0509 © 2013 The Authors. Published by Elsevier B.V.
Selection and peer-review under responsibility of KES International
doi:10.1016/j.procs.2013.09.207

randomness of any type of sequences by a single evaluation criteria. On the other hand, the RMT-test requires a long data sequences, which is often difficult in practical applications, since the lengths of real data do not meet the minimum length to satisfy $Q=L/N>1$, and $N>100$ required to justify the RMT formula.

The RMT-test has been applied on pseudo-random sequences and physical random sequences. The result shows that the degree of the randomness of output of those renowned random number generators are extremely high. The results were used to determine the criteria to define a good random numbers, in the former works by the authors [3-5]. Individual data sequence of 'good' random numbers show very little difference between the measured values of the 6-th moment and its ideal value derived from the RMT formula to be smaller than one percent. However, the sample mean over 100 samples lie around a few percent, due to large fluctuation.

The same tools was also applied on real-world data, such as log-return sequences of the pseudo-random numbers or the physical random numbers, to prove that the process of taking log-return adds a certain "off-randomness" to the data sequences thus lowers the level of randomness in such log-return series. Obviously, the degree of randomness of the log-return sequences is expected to be very low. The average difference between the measured moments and their ideal values become as large as 10-20 percent.

Due to the lack of existing good data having the intermediate level of randomness that satisfies the RMT conditions, in the real-world data, it has been difficult to fill the gap between 'good' random numbers and 'bad' random numbers. In order to pin-point the border between the 'highly random' data and the 'poorly random' data, a series of sequences in various levels of randomness are artificially prepared by shuffling in sufficient times until the sequences pass the NIST randomness test. The level of randomness that passes the NIST randomness test is supposed to be the border between the 'good random number' and the 'poor random number' in the sense described above.

2. Outline of the RMT-test

The method used in this paper is to compare the eigenvalue distribution of the correlation matrix, between N time series of length L, to the corresponding theoretical formula of the eigenvalue distribution derived from the random matrix theory in the limit of N and L going to infinity, keeping $Q = L/N$ as a constant[6-11]. This method is applied to the stock market in 2002 by Plerou [9]. The outline of the method is as below.

At first, a long sequence is cut into N pieces of equal length L, then shape them in an $L \times N$ matrix, by placing the first L elements in the 1st row of the matrix, and the next L elements in the 2nd row, etc., by discarding the remainder if the length of the sequence is not divisible by L. Then we normalize each column of the matrix to have zero mean and single variance, do the correlation matrix C by calculate the inner product of the normalized matrix. Obtain the eigenvalues of correlation matrix C by numerical calculation, measure the randomness of the sequence by comparing the eigenvalue distribution to the corresponding theoretical formula in Eq.(1). Because the maximum and minimum value of eigenvalue λ can be calculate by Q in Eq.(2), it is known that the parameter of Eq.(1) is Q only. As shown in Fig.1, if the two lines match, that data passes the RMT-test, and if they do not match, it fails the RMT-test. This is the qualitative evaluation of RMT-test.

$$P_{\mathrm{RMT}}(\lambda) = \frac{Q}{2\pi} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda} \tag{1}$$

$$\lambda_\pm = (1 \pm Q^{-1/2})^2 \tag{2}$$

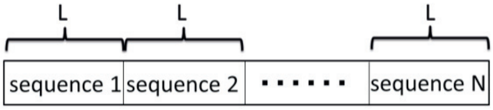

Fig.1 Method of dividing the random number sequence

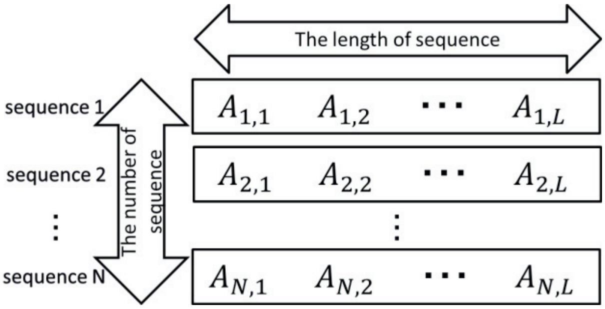
The length of sequence
sequence 1
sequence 2
sequence N
The number of sequence
$A_{1,1}$ $A_{1,2}$ ⋯ $A_{1,L}$
$A_{2,1}$ $A_{2,2}$ ⋯ $A_{2,L}$
$A_{N,1}$ $A_{N,2}$ ⋯ $A_{N,L}$

Fig.2 Arrangement of the data sequence of random numbers

Secondly, in order to compare the shape of the eigenvalue distribution which is so-called Marcenko-Pastur distribution we employ the moment method [4]. The quantitative evaluation based on the moment method compare the k-th moment of the obtained eigenvalues

$$m_k = \frac{1}{N}\sum_{i=1}^{N}\lambda_i^k \tag{3}$$

to the corresponding theoretical formula obtained from P_{RMT}

$$\mu_k = E(\lambda^k) = \int_{\lambda_-}^{\lambda_+}\lambda^k P_{RMT}(\lambda)d\lambda \tag{4}$$

The difference between m_k and $\mu_k$indicates the level of off-randomness of the data sequences. Using the 6th moment (*k=6*), the difference between the two quantities, which is called as the 'Error',

$$|\mathrm{Error}| = |m_6 - \mu_6| \tag{5}$$

has been obtained for massive data of pseudo-random sequences as well as physical random sequences [5].

The maximum Errors for both pseudo-random numbers, such as rand(), and the Mersenne Twister, are smaller than 1 percent for individual data sequence, and a few percent for the sample average plus twice the standard deviation over 100 samples. The results for the three different physical number generators are also in the same range. On the other hand, the Error for the log-return sequences of those random sequences ended up to be as large as 10-20 percent. Although it is expected that the borderline to divide the randomness and the off-randomness should reside between those two values, an independent source is required to determine the precise level of Error to be used as the randomness criterion.

3. NIST randomness test

An independent criterion to compare with the RMT-test would help to determine the objective border between the high randomness and low randomness. The NIST randomness test is one in the candidate who serves as such a role. The NIST SP 800-22 is the U.S. standard statistical randomness test whose source code is published on the website of NIST [2]. In addition, NIST randomness test is widely used as an assay whether this can be used as ciphers. The NIST SP 800-22 tests the randomness of data sequence in ASCII format consisting of 0 and 1 by means of applying the following 15 types of assays.

- Frequency (frequency of letters)
- Block Frequency (frequency of blocks)
- Cumulative Sums
- Runs
- Longest Run (within block)

- Rank (rank of matrix in binary)
- FFT(discrete Fourier transform)
- Non Overlapping Template (template matching without overlapping)
- Overlapping Template (template matching with overlapping)
- Universal (statistical assay by Maurer)
- Approximate Entropy
- Random Excursions
- Random Excursions Variant
- Serial
- Linear Complexity

The length of the sequence that NIST SP 800-22 test is recommended is 1,000,000 [12]. In addition, at least 55 samples are to be prepared in order to obtain statistically significant results.

4. Shuffled data having various levels of randomness

In exploring the criteria to be highly random by comparing the NIST randomness test and the RMT-test, it is essential to prepare the data sequences of a fixed length (very large, e.g., 1,000,000) in various levels of randomness. As a matter of fact, however, it is rather hard to find data sequences having intermediate levels of randomness. For example, the randomness levels of the output sequences of pseudo-random generators installed in computers are too high, and the randomness levels of the real-world data given by financial, social, or biological statistics are substantially low in general. Therefore, it is more convenient to create a set of artificial data in various levels of randomness by shuffling a regular sequence. As the number of shuffles increases, the randomness of the sequence is expected to grow. In order to meet the requirement of the NIST, 55 sequences are prepared each of which has the length of 1,000,000.

The initial sequence is constructed by placing fifty consecutive 0's and fifty consecutive 1's alternatively by ten thousand times to make the total length to be 1,000,000. This guarantees the equal frequencies of 0's and 1's in the data sequences. Then a shuffling is applied by exchanging randomly chosen two elements in the sequence. By repeating this shuffling T times, various data sequences are created. It is expected that the degree of randomness monotonically increases in proportional to T. For example, the sequences of 1 million shuffles, 2 million shuffles are constructed as follows, respectively.

(1) The initial sequence is shuffled by 1,000,000 times, and the result is written on the file named "the ransom number sequence of 1,000,000 shuffles".
(2) The sequence created in (1) is further shuffled by 1,000,000 times, and the result is written on the file named "the ransom number sequence of 2,000,000 shuffles".

By means of repeating this process, 55 samples in various levels of randomness are created for testing. The number 55 is required to apply the NIST test.

5. Results of the RMT- test and the NIST

First of all, it is necessary to confirm whether the randomness indeed increases according to the number of shuffles. The result of the quantitative evaluation of the RMT-test applied on the random number sequences created by shuffling 1,000,000~5,000,000 times is summarized in Fig. 3, where, the vertical axis shows the

|Error| in Eq. (5) averaged over 55 samples and the horizontal axis shows the number of shuffles, denoted by T. The moments of k=2,...,6 are shown, since the case k=1 has no error by definition.

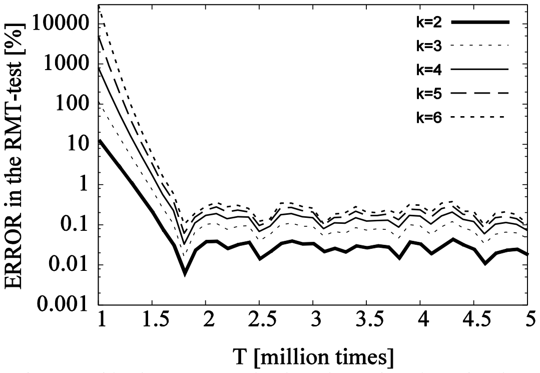

Fig.3 |ERROR| in the RMT-test is plotted as a function of T (k=2-6)

As the number of shuffles increases, the values of |Error| decreases monotonically before saturating at around T=1.8 million. This confirms that the shuffling indeed helps increasing the randomness of given sequences.

In order to cross check the result, the NIST test is applied on the same set of data. Since the NIST is designed to apply on a sequence in ASCII format consisting of 0 and 1, the data sequences are converted to the 0-1 data by using the median as the boundary. In all of the 15 assays in the NIST randomness test, the success or the failure is determined by the ratio of the number of samples that have passed the test. If the ratio is greater than or equal to a certain level, the sequence is determined to pass the test. The result of the NIST randomness test using the sequences of 1,000,000~5,000,000 shuffles is summarized in Table 1. The data used in the NIST randomness test are the same sequences used for the RMT-test. The result shows that the data with more than 1.7 million shuffles pass all the 15 assays, except for occasional failures at one particular assay.

Table 1. Result of the NIST randomness test

Time of shuffle (million times)	Pass rate	Time of shuffle (million times)	Pass rate	Time of shuffle (million times)	Pass rate
100	5/15	240	15/15	380	15/15
110	7/15	250	14/15	390	15/15
120	7/15	260	15/15	400	15/15
130	7/15	270	14/15	410	15/15
140	7/15	280	15/15	420	15/15
150	10/15	290	14/15	430	15/15
160	13/15	300	15/15	440	15/15
170	14/15	310	15/15	450	15/15
180	14/15	320	14/15	460	14/15
190	15/15	330	14/15	470	15/15
200	15/15	340	14/15	480	15/15
210	15/15	350	15/15	490	15/15
220	14/15	360	14/15	500	15/15
230	15/15	370	14/15		

Table 2. Comparison of the RMT-test and the NIST randomness test

RMT-test Error(%)	NIST randomness test Pass rate
29582.88	5/15
3803.87	7/15
572.09	7/15
101.80	7/15
22.27	7/15
5.79	10/15
1.60	13/15
0.60	14/15
0.38	14/15
0.36	15/15
0.35	15/15
0.34	15/15
0.34	14/15
0.32	15/15
0.31	15/15
0.30	14/15
0.29	15/15
0.28	14/15
0.28	15/15
0.28	14/15
0.26	15/15
0.25	15/15
0.24	14/15
0.22	14/15
0.21	14/15
0.21	15/15
0.21	14/15
0.20	14/15
0.20	15/15
0.20	15/15
0.19	15/15
0.19	15/15
0.18	15/15
0.18	15/15
0.18	15/15
0.14	15/15
0.12	14/15
0.11	15/15
0.10	15/15
0.10	15/15
0.10	15/15

The results of the RMT-test for k=2 and the NIST test using the same set of data are compared in Table 2, in which the data in Fig.3 is sorted in the descending order of |Error| to be compared to the results of the NIST test. Although the same values for the |Errors| in the RMT-test appear multiple times in Table 2, each value corresponds to a different value of shuffling time T.

The result in Table 2 shows that data which pass the NIST randomness test, by passing at least 14 out of 15 assays, are the sequences of |Error| < 0.60% in the RMT-test (shown in Table 2 in bold characters), the all sequences have a high pass rate because 14 out of 15 kinds or more are determined to pass.

6. Difference by the number of symbols

The purpose of this chapter is to check if the result depends on the number of symbols or not. Since the result obtained so far is on the case of the number of symbols being two (0 and 1), data using various kinds of symbols are used in order to check the symbol dependence.

The initial sequence is prepared by placing a hundred decimal numbers, 00, .., 99 in ascending order and combine ten thousand pieces of the same sequences to make the total length to be 1,000,000. By means of shuffling this initial sequence by T times, data sequences having various levels of randomness are created. The result using this decimal data is compared to the case of using two symbols (0 and 1) in Fig. 4.

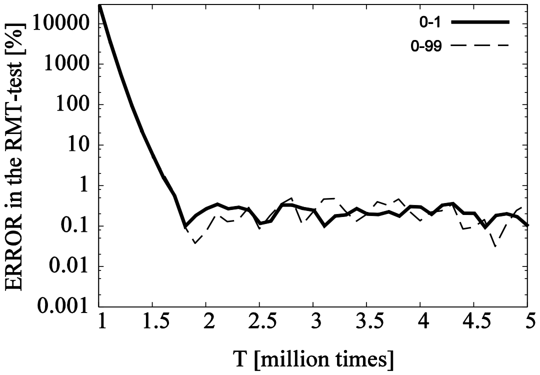

Fig.4 Comparison of two different numbers of symbols, 2 and 100 in the graphs of ERROR as a function of T (in million times).

The result in Fig.4 is summarized as follows:

1. |Error| decreases monotonically before the saturation point
2. |Error| fluctuates for less than 1% after reaching the saturation point

In addition, a comparison of the results of the RMT-test for the data composed of two symbols (0 and 1), ten symbols (0, 1, .., 9), and a hundred symbols (0,.., 99) are shown in Fig. 5. Likewise, the corresponding results of the NIST are shown in Fig.6, in which the vertical axis labelled by RESULT represents the number of passed assays out of total 15. As a result, different numbers of the symbols, 2, 10, and 100 did not show any significant difference, as shown in Figs. 5 and 6.

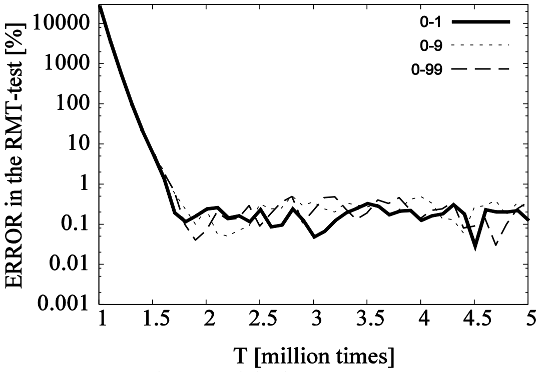

Fig.5 Result in the RMT-test

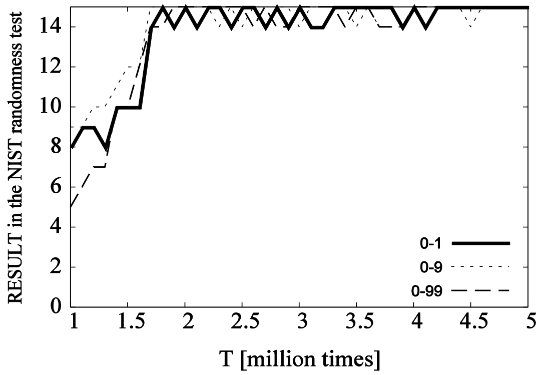

Fig.6 Result in the NIST randomness test

Finally, the cases of very high randomness given by the pseudo random generators and physical random numbers are examined for the sake of comparison. As shown in Table 3, the values of |Error| for all the cases are less than 0.3% in the RMT-test, and the rate of passed assays being 14 or larger out of 15 in the NIST.

For the sake of comparison, the degree of randomness of log-return sequences are examined as an example of low randomness, which are created by converting random number sequence $A_1, A_{2,} ..., A_N$ converted to $B_1, B_{2,} ..., B_{N-1}$ by using the following equation

$$B_i = \ln\left(\frac{A_i}{A_{i-1}}\right) \tag{6}$$

Table 3. Comparison using Pseudo-random number(LCG) and Three types of physical random number

Kind of random number	RMT-test Error(%)	NIST randomness test Pass rate
LCG	-0.2831	14/15
Hitachi	-0.1597	15/15
Toshiba	0.0026	15/15
Tokyo	-0.1194	15/15

Table 4. Comparison using log-return sequences as examples of low random data created by using Eq. (6).

Kind of random number	RMT-test Error(%)	NIST randomness test Pass rate
LCG	99.3042	5/15
Hitachi	98.8686	5/15
Toshiba	99.2463	5/15
Tokyo	98.7580	5/15

7. Conclusion and Discussions

In this article, the effect of shuffling has been confirmed both in the RMT-test and NIST randomness test, and the criteria of “The good random number in the case of considering NIST randomness test” has been determined to be |Error| < 0.60% in the RMT-test, that corresponds to the number of assays to be passed be at least 14 out of total 15 in the NIST. According to the survey, the sequences shuffled between 1.7 million and 1.8 million times failed only in the assay of “Runs”, while the sequences more than 1.8 million times failed only in the assay of “Non Overlapping Template”. This fact seems to imply that the randomness calculated by the RMT-test is not related to the assay of “Non Overlapping Template” in the NIST test.

The criterion of randomness |Error| < 0.60% reported in this article is substantially small compared to the value |Error| < 5% obtained in our previous result. In addition, if the number of shuffle is higher, the pass rate in NIST randomness test is high. In fact, NIST randomness test is widely used as the assay by the cipher. Therefore, the sequence is suitable as a cipher when randomness is high by the result in 5.3. From the above, the sequences that their pass rate of NIST randomness test are 14/15 or 15/15 and their values of the error calculated by RMT-test are 0.60% or less are determined as “the good random number appropriate degree as a cipher”.

References

[1] Park, S. and Miller, K.: Random Number Generators: Good Ones are Hard to Find, *Communication of ACM*, 1988; **31**: pp.1192-120.

[2] NIST: http://csrc.nist.gov/groups/ST/toolkit/rng/documentation software.html

[3]Yang, X., Itoi, R., and Tanaka-Yamawaki, M.: Testing Randomness by Means of RMT Formula, *Intelligent Decision Technologies, SIST,* 2011; **10**: pp.589-596.

[4]Yang, X., Itoi, R., and Tanaka-Yamawaki, M.: Testing randomness by means of Random Matrix Theory, *Progress of Theoretical Physics Supplement,* 2012; **194**: pp. 73-83.

[5] Tanaka-Yamawaki, M., Yang, X., and Itoi, R.: Moment Approach for quantitative evaluation of randomness based on RMT formula, *Intelligent Decision Thechnologies*.:*SIST (Springer),* 2012; **16**: pp. 423-432.

[6] Wigner, E. P.: Ann. Math., 1958; **67**: pp. 325-327.

[7] Laloux, L., Cizeaux, P., Bouchaud, J., and Potters, M.: Noise Dressing of Financial Correlation Matrices, *Physical Review Letters*, 1998; **83**: pp.1467-1470.

[8] Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L. A. N., and Stanley, H. E.: *Physical Review Letters*, 1999; **83**: pp.1471-1474.

[9] Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L. A. N., and Stanley, H. E.: Random Matrix Approach to Cross Correlation in Financial Data, *Physical Review E*, 2002; **65**: no.066126.

[10] Tanaka-Yamawaki, M.: Cross correlation of intra-day stock prices in comparison to Random Matrix Theory, *Intelligent Information Management,* 2011; **3**: pp.65-70.

[11] Tanaka-Yamawaki, M.: Extracting Quarterly Trends of Tokyo Stock Market by Means of RMT-PCA, *Advances in Knowledge-Based and Intelligent Information& Engineering Systems*, 2012; DOI: 10.3233/978-1-61499-1-05-2-2028: pp. 2028-2036.

[12] Rukhin, A. Soto, J., Neckvatal, J., Smid, M., Barker, E., Leigh, S., Levenson, M., Vangel, M., Banks, D., Heckert, A., Dray, J., and Vo, S.: A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications, *NIST(National Institute of Standards and Technology, U.S. Department of Commerce)*, 2010; Special Publication 800-22, Revision 1a: pp. 1-8.