Computer Science Journal of Moldova, vol.**7**, no.3(21), 1999

Definition of necessary length of random sequences at execution of the empirical tests (criterion χ^2)

O.M. Petrova

Abstract

This article is devoted to the problem solution, arising from the necessity to test a sequence of random numbers. The application of Pearson's criterion imposes restriction from below on the number of elements of the random sequence. In the article, the set of the classical empirical tests is considered and the expressions for calculation of necessary length of the random sequence are given.

1 Introduction

Random numbers are used by a person for solving problems, very various both on subjects, and their significance. It comprises modeling the various natural and socioeconomic phenomena, problem of the numerical analysis, check of efficiency of analyzed computing algorithms. Besides, the rapid development of computer techniques has resulted in appearance of the whole industry of cyber-amusements, and the computer games almost always implement "random" situations; the organization of advertising on the Internet through the banner-exchange system also provides the use of random sequences. Another sphere of application of the random numbers is provision of information security (commercial, official and military) using cryptographic algorithms. The system of the digital signature based on the DSS standard is frequently applied for the data protection and authentication. When implementing the DSA algorithm (the kernel of DSS), the generation of

the random numbers for forming users' private keys and secret number for each message when signing it, is necessary.

The devises using random processes in nature, for example the white noise, or generators of pseudo-random numbers, functioning on the basis of the mathematical tool (linear congruent generator, lagged-Fibonacci generator etc.) can serve as a source of the random numbers.

2 Problem statement

In the presence of the sequence of pseudo-random numbers, there is a problem to determine the degree of randomness of this sequence. The statistical theory gives some quantitative criterions of randomness, from which most widespread is the criterion χ^2. The application of this criterion is stipulated by the fact, that random numbers, frequently used in practice, belong to the set of positive integers and the number of their variants is finite. It is the constituent of the empirical tests and allows determining the measure of deviation of empirical distribution from hypothetical one on the formula offered by Pearson:

$$V = \sum_{1 \le s \le k} \frac{(Y_s - n \cdot p_s)^2}{n \cdot p_s} \tag{1}$$

where

Y_s – number of tests, which have fallen into the s category;

p_s – probability of test results' falling into the s category;

k – number of categories;

n – total quantity of tests.

According to the meaning χ^2 and number of degrees of freedom $\nu = k - 1$, the probability of an event occurrence (i.e. correctness of the expressed hypothesis) is calculated. The number of tests n should be such, that for any p_s the product $n \cdot p_s$ was not less than 5. Therefore, the problem to determine the least meaning of number of tests (or the number of random sequence elements), necessary for fulfillment of the condition, is set:

$$\forall p_s : n \cdot p_s \ge 5$$

Or

$$\forall s: \ \min(n \cdot p_s) = n \cdot \min(p_s) \geq 5$$

This implies, that the minimum meaning of necessary number of elements of the random sequence should be determined by the formula:

$$n_{\min} \geq 5/\min(p_s), \qquad 1 \leq s \leq k \tag{2}$$

As the empirical tests [1, 2], considered below, manipulate with groups of elements of the sequence, the number of categories k and the meaning of probability p_s for each test will be individual. Thus, the initial problem is reduced to definition $n_{\min i}$ for each (i^{th}) test separately and further:

$$n_{\min} \geq \max_i(n_{\min i}) \tag{3}$$

3 Necessary number of trials for the classical empirical tests

Frequency Test

The random numbers of a sequence should be uniformly distributed between 0 and $d-1$ (d – basis of number system, at binary notation the meaning d is determined by the number of binary symbols that are included in a base sequent). During testing, it is calculated how many times each number was met in the sequence. The Pearson's criterion with $k = d$ and probabilities $p_s = 1/d$ is applied further. Thus

$$n_1 \geq 5/p_s = 5 \cdot d$$

Serial Test

The test checks the degree of randomness and independence of the sequences of m of random numbers following one after the other. For this purpose, the quantity of each sequence met is calculated. The Pearson's criterion with $k = d^m$ and probabilities $p_s = 1/d^m$ is further applied. In this case

$$n_2 \geq 5 \cdot m/p_s = 5 \cdot m \cdot d^m$$

Overlapping M - Tuple Test
This test is a modification of the serial test. The initial sequence of random numbers is divided into groups of k_1 elements each. After that the test analyses the sequences from first k_2 (statistics V_{k2}) and all k_1 (statistics V_{k1}) elements of the group [3]. As the calculation is executed separately on each statistics, taking into account, that $k_1 > k_2$, we obtain

$$n_3 \geq 5/p_{k_1} = 5 \cdot d^{k_1}$$

Permutation Test
The initial sequence is divided into the groups of t of the following successively one after the other elements each. In each group it is possible $t!$ variants of a relative disposition of numbers altogether. It is calculated, how many times each concrete disposition has been met, then the Pearson's criterion with $k = t!$ and probabilities $p_s = 1/t!$ is applied. As only those subsequences are considered, where all the elements are different and the probability of a similar situation is calculated by the formula

$$p_{\neq} = \prod_{i=0}^{t-1}(d-i)/d^t,$$

then

$$n_4 \geq \frac{5 \cdot t}{p_s \cdot p_{\neq}} = 5 \cdot t \cdot t! \cdot d^t / \prod_{i=0}^{t-1}(d-i)$$

The "Maximum of T" Test
The algorithm of the "maximum of t" test that is also some modification of the serial test, consists in the following. On the first step, the initial sequence of random numbers is divided into the groups of t elements, following one after the other, and the maximum element is singled out in each group. On completion of the procedure we have a new sequence of "maximums of t". After this, applying the statistical criterion, we determine, whether the frequency of appearance of each "maximums of t" sequence's member corresponds to specific probability. The problems connected to implementation of the given test are

minutely considered in [4, 5]. For the classical variant of the test, the necessary number of elements of the random sequence should be:

$$n'_5 \geq 5 \cdot t \cdot d^t$$

For the simplified version of the test, when the initial items of the sum up to $s = i_{start}$ are excluded, in the formula (1) on respective reduction of degrees of freedom number, we have

$$n''_5 \geq 5 \cdot t \cdot \left[\left(\frac{i_{start} + 1}{d} \right)^t - \left(\frac{i_{start}}{d} \right)^t \right].$$

The Poker - Test

The classical poker – test considers N groups of five following one after the other random numbers. In each group, the calculation of unmatched numbers is executed. The number of categories, at the same time, is equal to 5:

- 5 different ones – all of them are different;
- 4 different ones – one pair;
- 3 different ones – two pairs or 3 of one kind;
- 2 different ones – full collection or 4 of one kind;
- no different ones – five of one kind.

In the test the number of groups is calculated, in which there are $5 - 4 - 3 - 2 - 0$ of different numbers and then the χ^2 criterion is applied.

Some modification of the test consists in the following: in each from m groups of the following one after the other four random numbers, the number of coincident pairs is determined – no coincidences, one pair, two, three or four. Then it is calculated, how many groups correspond to each category ($0-1-2-3-4$ pairs), then the χ^2 criterion is applied. In this case, the minimal meaning p_s from the formula (2) will be:

$$p_{\min} = d/d^4 = 1/d^3.$$

Thus, the expression for n_6 will become:

$$n_6 \geq 5 \cdot 4/p_{\min} = 20 \cdot d^3$$

Gap Test
The length of intervals between appearances of meanings belonging to some specific segment $[\alpha, \beta]$, is checked up in this test. For this purpose the number of length intervals $0, 1, 2, \ldots, (t-1)$ and full number of intervals of larger length ($\geq t$) are determined. The obtained meanings are treated using Pearson's criterion with $k = t + 1$ and probabilities $p_0 = p$, $p_1 = p(1-p)$, $p_2 = p(1-p)^2, \ldots, p_t = p(1-p)^t$, where $p = (\beta - \alpha)/d$. Meanings of the series members $p_0, p_1, p_2, \ldots, p_t$ monotonically decrease. Hence, $p_{\min} = p_t$. In this case, the desired expression will become:

$$n_7 \geq 5 \cdot \sum_i p_i \cdot i / \left(p \cdot (1-p)^t\right), \qquad p_i = p \cdot (1-p)^i.$$

It is necessary to mark that the "gap test" is the only one, where the decision of the delivered problem does not depend on the base number. The result only depends on a relative breadth of an interval.

Runs Up Test
This test analyses length (from 1 up to t) of monotonous non-decreasing subsequences in the initial sequence of random numbers. For the given test the desired value is determined by the formula:

$$n_8 \geq 5 \cdot \sum_i p_i \cdot i / p_{\min} = 5 \cdot t! \cdot \sum_i p_i \cdot i, \qquad p_i = i/(i+1)!.$$

Coupon Collector Test
In the initial sequence of random numbers, the lengths of sequential segments (from 1 up to t), necessary to collect "totality" of integers from 0 up to $w-1$, $w < d$ are determined. The given algorithm calculates how many times each length of segment was met and applies Pearson's criterion. Probability of appearance of this or that size of segment is the following:

$$p_r = \frac{d!}{w! \cdot d} \cdot \sum_{j=0}^{w-1} \frac{(-1)^j \cdot \left(\frac{w-j-1}{d}\right)^{r-1}}{(w-j-1)! \cdot j!}.$$

Considering p_r function as a dependence from r, (in the presence of sufficient choice t), its maximum is on a section from w up to t. Hence, it is necessary to search for minimum meanings of probability on the section ends:

$$p_{\min} = \min \left\{ \frac{(d-1)!}{w!} \times \right.$$

$$\left. \times \min \left(\sum_{j=0}^{w-1} \frac{(-1)^j \cdot \left(\frac{w-j-1}{d}\right)^{w-1}}{(w-j-1)! \cdot j!}, \sum_{j=0}^{w-1} \frac{(-1)^j \cdot \left(\frac{w-j-1}{d}\right)^{t-1}}{(w-j-1)! \cdot j!} \right), \left(1 - \sum_{i=w}^{t} p_i\right) \right\}.$$

Then necessary length of a random sequence should be the following:

$$n_9 \geq 5 \cdot \sum_r p_r \cdot r / p_{\min}.$$

To pass all the above-mentioned empirical tests, the random sequence length should be not less than n:

$$n \geq \max_{1 \leq j \leq 9} (n_j).$$

As the value n_1 is admittedly less than n_2, it can be excluded from consideration[1]. Similarly, if, when testing using lapped m-tuple test, $k_1 = 3$, then n_3 will be less than n_6, and n_3 will not influence the final meaning of n.

As an example we shall consider a sequence of random decimal numbers ($d = 10$), analyzed using empirical tests [6]. In this case $n_1 > 50$, $n_2 > 200000$, $n_3 > 5000$, $n_4 > 9920$, $n_5 > 200000$ (classical test and $n_5 > 979$ for the reductive variant), $n_6 > 20000$, $n_7 > 7976$, $n_8 > 6185$, $n_9 > 320680$. So, in order to the sequence pass all the tests, its length should be not less than 320680. This is the necessary condition, but not sufficient (in practice, the number of the sequence elements should constitute 400000 at least, that was confirmed by experiments).

[1]Non-fulfillment of the Pearson's condition when executing the frequency test testifies that the given sequence will not pass any of the described tests.

4 Conclusion

The obtained expressions allow determining necessary amount of experiments for getting reliable estimate of the random sequence prior to the beginning of testing. Besides, using them, it is possible to decide the inverse problem as well – if there is a certain sequence of random numbers, it becomes possible to determine a concrete set of the empirical tests for its analysis, in which the Pearson's criterion is executed. If it is necessary to pass all the tests for the given random sequence, and the number of elements is not sufficient, or there are restrictions due to the hardware (for instance, memory volume), it is possible to get the desirable result by selection of parameters. Simultaneously, one should bear in mind, that the more the length of the analyzed sequence, the higher the reliability of χ^2 criterion and the obtained estimate is only correct in the volume of the executed tests.

When the increased requirements are presented to random numbers, and it is especially typical for cryptographic systems of information security, the most careful analysis of random sequences using the statistical tests is necessary [6]. The more tests will successfully pass a researched sequence, the more reliable will be an outcome.

References

[1] Knuth Donald E. *The art of computer programming,*, V. 2 Seminumerical Algoritms. - M.: Mir, 1977. p. 727. (In Russian)

[2] Prohorov Yu.V. (Editor-in-Chief), *Mathematic Encyclopedia Dictionary* - M.: Sov. Encyclopedia, 1988. p. 847. (In Russian)

[3] Wegentkittl S., *Empirical testing of pseudorandom number generator* Master's thesis, University of Salzburg, Austria, 1995.

[4] Shapira A., *The Discrete Runs Test and the Discrete Maximum of t Test.*, Technical Report CS 96-15, ESCE Department Rensselaer Polytehnic Institute, 1996.

[5] Oleinik W.L., Petrova O.M. *Testing of random sequences*, Acta Academia, Kishinev, 1999. pp. 127 - 134. (In russian)

[6] *The package of applied programs for random numbers sequences testing (DEKART Random Run's Tests).*, Programmer manual. Operator manual. "Dekart S.R.L.", 1998. (In russian)

O.M.Petrova,
"Dekart S.R.L"
51a, Alexandru cel Bun, Kishinev,
MD2012, MOLDOVA
phone: (373+2) 24–55–80
e-mail: *petrova@dekart.com*

Received November 1, 1999