A new statistical testing for random numbers and its application to some cryptographic problems.

B.Ya. Ryabko, V.S. Stognienko, and Yu.I. Shokin[1]
Institute of Computational Technology of Siberian Branch of Russian Academy of Science
{ryabko, vss, shokin}@adm.ict.nsc.ru

We address the problem of testing the hypothesis H_0 that the letters from some alphabet $A = \{a_1, a_2, \ldots, a_K\}$ are distributed uniformly (i.e. $p(a_1) = p(a_2) = \ldots = p(a_k) = 1/K$) against the alternative hypothesis H_1 that the true distribution is not uniform, in the case when K is large.

First, let us explain the main idea of the new test. Let there be given a sample which can be used for testing. The sample is divided into two parts, which are called the *training sample* and the *testing sample.* The training sample is used for estimation of frequencies of the letter occurrences. After that the letters of the alphabet A are combined into subsets $A_1, A_2, \ldots, A_s, s \geq 2$, in such a way that, first, one subset contains letters with close (or even equal) frequencies of occurrence and, second, s is much less than K (say, $K = 2^{20}, s = 2$). Then, the set of subsets $\{A_1, A_2, \ldots, A_s\}$ is considered as a new alphabet and the new hypotheses $\hat{H}_0 : p(A_1) = |A_1|/K, p(A_2) = |A_2|/K, \ldots, p(A_s) = |A_s|/K$ and $\hat{H}_1 = \neg\hat{H}_0$ are tested based on the second ('testing') part of the sample. Obviously, if H_0 is true, then $\hat{H}_0$ is also true and, if $\hat{H}_1$ is true, then H_1 is true. That is why this new test can be used for testing the initial H_0 and H_1. The idea of such a scheme is quite simple. If H_1 is true, then there are letters with relatively large and relatively small probabilities. Generally speaking, the high-probable letters will have relatively large frequencies of occurrence and will be accumulated in some subsets A_i whereas low-probable letters will be accumulated in the other subsets. That is why this difference can be found based on the testing sample. It should be pointed out that a decrease in the number of categories from large K to small s can essentially increase the power of the test and, therefore, can essentially decrease the required sample size. For example, let the number of categories K is even and let H_1 is as follows: $p^1_{i_1} = p^1_{i_2} = \ldots = p^1_{i_{K/2}} = (1 + \delta)/K$, $p^1_{i_{(K/2)+1}} = \ldots = p^1_{i_K} = (1 - \delta)/K$, where $\delta \in (0, 1)$, $\{i_1, \ldots, i_{K/2}\} \bigcup \{i_{(K/2)+1}, \ldots, i_K\} = \{1, \ldots, K\}$. It turns out that the new test can be successfully applied when the total sample size is $O(\sqrt{K})$ instead of *const* K for usual χ^2 test.

Claim 1. *Let the new test be applied for testing H_0 and H_1. Then, for each $\delta \in (0, 1)$ and $\alpha \in (0, 1)$ there exist such a training sample size $m(\delta)$ and a testing sample size $n(\delta)$ that i) $(m(\delta) + n(\delta)) = O(\sqrt{K})$, ii) the level of significance of the test is α and the Type II error is less than 1/2.*

The natural question is why the case of large alphabet size K can be interesting. It turns out that there exist such random bit sequences $x_1x_2\ldots$ which, from the one hand, are very far from truly random and, on the other hand, the distribution of the words $x_1 \ldots x_s$, $x_{s+1} \ldots x_{2s}$, $x_{2s+1} \ldots x_{3s}, \ldots$ is uniform for relatively large s. (We call such processes *two-faced.*) Hence, if a test is designed for two-faced processes testing, it must deal with the relatively large word length. Therefore, the alphabet contains all s-bit words and the alphabet size $K (= 2^s)$ grows exponentially when s grows. It worth noting that two-faced sequences are often met in cryptography. For example, in fact, pseudorandom number generators are designed to generate such sequences [1, 2].

Claim 2. *For each integer $k \geq 1$ and $\varepsilon \in (0, 1)$ there exist such processes $T(k)$ and $T^*(k)$ that the k-order Shannon entropy (h_k) of the processes $T(k)$ and $T^*(k)$ equals 1 bit per letter whereas the limit Shannon entropy (h_∞) equals ε.*

The processes $T(k)$ and $T^*(k)$ are Markov chains of connectivity (memory) k, which generate letters from $\{0, 1\}$. Let π be such that $-(\pi \log_2 \pi + (1 - \pi) \log_2(1 - \pi)) = \varepsilon$. The process $T(1)$ is defined by conditional probabilities $P_{T(1)}(0/0) = \pi, P_{T(1)}(0/1) = 1 - \pi$ (obviously, $P_{T(1)}(1/0) = 1 - \pi, P_{T(1)}(1/1) = \pi$) and the process $T^*(1)$ is defined by $P_{T^*(1)}(0/0) = 1 - \pi, P_{T^*(1)}(0/1) = \pi$. Assume that $T(k)$ and $T^*(k)$ are defined and describe $T(k+1)$ and $T^*(k+1)$ as follows $P_{T(k+1)}(0/0u) = P_{T(k)}(0/u)$, $P_{T(k+1)}(0/1u) = P_{T^*(k)}(0/u)$, $P_{T^*(k+1)}(0/0u) = P_{T^*(k)}(0/u)$, $P_{T^*(k+1)}(0/1u) = P_{T(k)}(0/u)$, for each $u \in \{0, 1\}^k$.

We applied the adaptive chi-square test to ciphered English and Russian texts. Apparently, the problem of constructing tests which can distinguish ciphered texts from random sequences can be considered as a good example for estimation of a power of statistical tests, because, on the one hand, it is known that ciphered texts cannot be completely random in principle and, on the other hand, the ciphers are constructed in such a way that the ciphered sequences should look random. (As much as possible). Besides, this problem is of some interest for cryptography [2].

The texts were combined in large files and each file was ciphered by Rijndael (AES) with 128-bit block length in such a way that one key was used for ciphering of all blocks from one file. Then we took 40 files of texts in English and 40 such files in Russian. Each ciphered file was tested for randomness using the new algorithm and the usual chi- square test. The power of the new test was larger than of the usual chi- square test, in spite of the fact that the maximal number of rejection was taken, when the usual chi- square test was applied (the maximum was taken over all possible block lengths). When both tests were applied for testing ciphered 512000− byte files in English, the hypothesis H_0 was rejected 28 times by the new test and 8 times by the usual chi- square test. For Russian texts the numbers of rejections were 24 and 5, correspondingly.

References

[1] Maurer U. *A universal statistical test for random bit generators.*, Journal of Cryptology, v.5, n.2, 1992, pp.89-105.

[2] Schneier B. *Applied Cryptography.* Wiley, 1996.

[1] The part of this work was done when B. Ryabko visited Centre of Excellence of European Union, Institute of Mathematics of Polish Academy of Science, Warsaw. B.Ryabko was also supported by INTAS under Grant no. 00-738.