Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Statistical Planning and Inference 133 (2005) 95–110
journal of statistical planning and inference
www.elsevier.com/locate/jspi

Using information theory approach to randomness testing☆

B.Ya. Ryabko*, V.A. Monarev

Department of Applied Mathematics & Cybernatics, Siberian State University of Telecommunication & Computer Science, Kirov str. 86, Novosibirsk 630102, Russian Federation

Received 1 July 2003; accepted 20 February 2004
Available online 1 July 2004

Abstract

We address the problem of detecting deviations of binary sequence from randomness,which is very important for random number (RNG) and pseudorandom number generators (PRNG). Namely, we consider a null hypothesis H_0 that a given bit sequence is generated by Bernoulli source with equal probabilities of 0 and 1 and the alternative hypothesis H_1 that the sequence is generated by a stationary and ergodic source which differs from the source under H_0. We show that data compression methods can be used as a basis for such testing and describe two new tests for randomness, which are based on ideas of universal coding. Known statistical tests and suggested ones are applied for testing PRNGs. Those experiments show that the power of the new tests is greater than of many known algorithms.

MSC: 62B10; 62G10; 62M07; 62M10; 94A29

Keywords: Hypothesis testing; Randomness testing; Random number testing; Universal code; Information Theory; Random number generator; Shannon entropy

1. Introduction

The randomness testing of random number and pseudorandom number generators is used for many purposes including cryptographic, modeling and simulation applications

☆ Supported by INTAS Grant No. 00-738 and Russian Foundation for Basic Research under Grant No. 03-01-00495.

* Corresponding author. Tel.: +7-3832-284938; fax: +1-3832-668030.
E-mail address: ryabko@adm.ict.nsc.ru (B.Ya. Ryabko)

0378-3758/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2004.02.010

(see, for example, Knuth, 1981; L'Ecuyer, 1994; Maurer, 1992; Mezenes et al., 1996). For such applications a required bit sequence should be true random, i.e., by definition, such a sequence could be interpreted as the result of the flips of a "fair" coin with sides that are labeled "0" and "1" (for short, it is called a random sequence; see Rukhin et al., 2001). More formally, we will consider the main hypothesis H_0 that a bit sequence is generated by the Bernoulli source with equal probabilities of 0's and 1's. Associated with this null hypothesis is the alternative hypothesis H_1 that the sequence is generated by a stationary and ergodic source which generates letters from $\{0, 1\}$ and differs from the source under H_0.

In this paper we will consider some tests which are based on results and ideas of Information Theory and, in particular, the source coding theory. First, we show that a universal code can be used for randomness testing. (Let us recall that, by definition, the universal code can compress a sequence asymptotically till the Shannon entropy per letter when the sequence is generated by a stationary and ergodic source.) If we take into account that the Shannon per-bit entropy is maximal (1 bit) if H_0 is true and is less than 1 if H_1 is true (Billingsley, 1965; Gallager, 1968), we see that it is natural to use this property and universal codes for randomness testing because, in principle, such a test can distinguish each deviation from randomness, which can be described in a framework of the stationary and ergodic source model. Loosely speaking, the test rejects H_0 if a binary sequence can be compressed by a considered universal code (or a data compression method).

It should be noted that the idea to use the compressibility as a measure of randomness has a long history in mathematics. The point is that, on the one hand, the problem of randomness testing is quite important for practice, but, on the other hand, this problem is closely connected with such deep theoretical issues as the definition of randomness, the logical basis of probability theory, randomness and complexity, etc. (see Kolmogorov, 1965; Li and Vitanyi, 1997; Knuth, 1981; Maurer, 1992). Thus, Kolmogorov suggested to define the randomness of a sequence, informally, as the length of the shortest program, which can create the sequence (if one of the universal Turing machines is used as a computer). So, loosely speaking, the randomness (or Kolmogorov complexity) of the finite sequence is equal to its shortest description. It is known that the Kolmogorov complexity is not computable and, therefore, cannot be used for randomness testing. On the other hand, each lossless data compression code can be considered as a method for upper bounding the Kolmogorov complexity. Indeed, if x is a binary word, ϕ is a data compression code and $\phi(x)$ is the codeword of x, then the length of the codeword $|\phi(x)|$ is the upper bound for the Kolmogorov complexity of the word x. So, again we see that the codeword length of the lossless data compression method can be used for randomness testing.

In this paper we suggest tests for randomness, which are based on results and ideas of the source coding theory.

Firstly, we show how to build a test basing on any data compression method and give some examples of application of such test to PRNG's testing. It should be noted that data compression methods were considered as a basis for randomness testing in literature. For example, Maurer's Universal Statistical Test, Lempel–Ziv Compression Test and Approximate Entropy Test are connected with universal codes and are quite popular in practice (see, for example, Rukhin et al., 2001). In contrast to known methods, the suggested approach