STATISTICAL TESTING of RANDOMNESS: NEW and OLD PROCEDURES

appeared as Chapter 3 in *Randomness through Computation*, H. Zenil ed. World Scientific, 2011, 33-51

Andrew L. Rukhin
Statistical Engineering Division
National Institute of Standards and Technology
Gaithersburg, MD 20899-0001 USA

1 Why I was drawn in the study of randomness testing: introduction

The study of randomness testing discussed in this chapter was motivated by attempts to assess the quality of different random number generators which have widespread use in encryption, scientific and applied computing, information transmission, engineering, and finance. The evaluation of the random nature of outputs produced by various generators has became vital for the communications and banking industries where digital signatures and key management are crucial for information processing and computer security.

A number of classical empirical tests of randomness are reviewed in Knuth (1998). However, most of these tests may pass patently nonrandom sequences. The popular battery of tests for randomness, Diehard (Marsaglia 1996), demands fairly long strings (2^{24} bits). A commercial product, called CRYPT-X, (Gustafson et al. 1994) includes some of tests for randomness. L'Ecuyer and Simard (2007) provide a suite of tests for the uniform (continuous) distribution.

The Computer Security Division of the National Institute of Standards and Technology (NIST) initiated a study to assess the quality of different random number generators. The goal was to develop a novel battery of stringent procedures. The resulting suite (Rukhin et al, 2000) was successfully applied to pseudo-random binary sequences produced by current generators. This collection of tests was not designed to identify the best possible generator, but rather to provide a user with a characteristic that allows one to make an informed decision about the source. The key selection criteria for inclusion in the suite were that the test states its result as a numerical value (P-value) indicating "the degree of randomness", that the mathematics behind the test be applicable in the finite sequence domain, and that there be no duplication among the tests in the suite. All of the tests are applicable for a wide range of binary strings size and thus exhibit considerable flexibility. While an attempt was made to employ only procedures which are optimal from the point of view of statistical theory, this concern was secondary to practical considerations.

In the next sections we review some of the tests designed for this purpose. Most of them are based on known results of probability theory and information theory, a few of these procedures are new. Before doing this, however, we discuss one of the first applications of the test suite.

1.1 Testing block ciphers: statistical theory and common sense

One application of the tests of randomness is block ciphers. These ciphers are widely used in cryptographic applications. Ten years ago NIST carried out a competition for the development of the "Advanced Encryption Standard (AES)". Its goal was to find a new block cipher which could be used as a standard. One of the requirements was that its output sequence should look like a random string even when the input is not random.

Indeed, one of the basic tests for the fifteen AES candidates was "Randomness Testing of the AES Candidate Algorithms," whose aim was to evaluate these candidates by their performance as random number generators (Soto and Bassham, 2001). The winner of the competition, the Rijndael algorithm, as well as other finalists, Mars, RC6, Serpent and Twofish, were used in the experiment involving randomness testing of their bit output by using statistical procedures in the NIST test suite.

To measure their performance, a numerical characteristic of the degree of randomness was required. For a given test, such a characteristic is provided by the P-value which quantifies the likelihood of a particular, observed data sequence under the randomness assumption. We discuss the P-values, their interpretation, and the difficulties of assigning them in section 1.2. Meantime, it should be mentioned that although different from the probability of the randomness hypothesis being true, P-values bear the same interpretation: small P-values indicate a disagreement between the hypothesis and the observed data. Although larger P-values do not imply validity of the hypothesis at hand, when a string is being tested by a number of tests, they are necessary to continue the study of multifaceted aspects of non-randomness. A very important property of P-values is that they are uniformly distributed over the unit interval when the tested null hypothesis is correct. Thus, about 10 P-values should be expected in the interval $(0, 0.01)$ if the true hypothesis is being tested $1,000$ times.

In a simplified description, in the AES evaluation experiment each algorithm generated a long string (about 2^{20} bits) stemming from the data type and the keysize Altogether 300 different data sequences were generated by each algorithm under combinations of data type and keysizes. Each of these sequences was subject to a randomness test from the suite resulting in a pass/fail decision. This decision was based on comparison of the P-value and a significance level which was chosen as 0.01. The P-values obtained were tested for uniformity. The sum of overall number of pass/fail decisions over all sequences was used as the statistic to judge randomness according to the particular test: if this sum is below a certain bound, the data set is deemed to have passed this statistical test. Thus, each algorithm under a particular test of randomness generated three hundred decisions with regard to agreement of the output and the randomness hypothesis. Both the sum of pass/fail decisions and the characteristic of uniformity (based on χ^2-test discussed in section 1.2) were used in the final assessment. If none of the P-values fell below 0.0001, the sample was believed to have satisfied the criterion for being random from the point of view of the given statistical test.

This procedure has been criticized (Murphy, 2000) from several perspectives. According to principles of statistical inference it is preferable to have one summary statistic on the basis of a long sequence rather than a number of such statistics obtained from shorter subsequences. But testing of encryption algorithms is not a purely statistical exercise. The validation of uniform P-values does not enter into the calculation of the power of a particular test, yet it can be seriously recommended from a practical point of view. The whole problem of testing randomness is not as unambiguous as the parametric hypothesis testing problems of classical mathematical statistics.

The same common sense approach led to concatenation of cipherblocks derived from random keys and different plaintexts. Murphy (2000) compares this process to interleaving or decimating an underlying message, so that either some signal is removed or some noise is added. Exploration and validation of various transmission regimes may be a more apt comparison. Besides, from the cryptographic point of view, the entropy of the concatenated text is larger than that of the original sequence. The concept of *randomization* in statistical design theory presents a similar idea. Randomized designs do not lead to better performance of statistical procedures if the postulated model is correct. However, they provide a commonly recommended safeguard against violations of the model assumptions.

Equally misguided seem to be arguments in favor of tests which are invariant to data transformations. In the statistical inference context this principle is violated by virtually all proper Bayes procedures. In AES testing context symmetrization over all data permutations is unpractical if not impossible. The use of random plaintext/random 128-bit key data type employed in the preliminary testing of AES candidates resulted in a mere permutation of the plaintext blocks. This data category was abandoned at later stages.

The concept of admissibility of tests (cf. Rukhin, 1995) was not very helpful when testing randomness. Indeed, practical considerations led to inclusion in the suite not only the frequency (monobit or balance) test, but also the frequency test within a block. From the theoretical point of view the latter is superfluous, from the pragmatic perspective it is a useful check-up.

1.2 P-values, one-sided alternatives versus two-sided alternatives and χ^2-tests

One of the principal difficulties of studying tests of randomness in statistical hypothesis formulation is that the null hypothesis, according to which the observed bits represent independent Bernoulli random variables with the probability of success 1/2, is typically false. Indeed, this is certainly the case in the previous example of block ciphers, and more generally for all pseudorandom number generators which are based on recursive formulas. In view of this fact, one may expect only a measure of randomness to be attested to by a given string.

To explain, we recall the basic definitions of the classical hypothesis testing which traditionally involve the so-called parameter space Θ. This set indexes all possible probability distributions of the observed data. A subset Θ_0 of Θ corresponding to special parametric values of interest is the null hypothesis $H_0 : \theta \in \Theta_0$. In many problems one can specify the alternative hypothesis $H_1 : \theta \in \Theta_1$, often taking by the default $\Theta_1 = \Theta - \Theta_0$. However this specification is not straightforward and this is the case of randomness testing. For example, if Θ is formed by real numbers and the null hypothesis specifies a particular probability distribution for the data, the alternative could be all probability distributions different from that one, or perhaps all probability distributions which are stochastically larger or smaller. (Think of the life-time distribution of a device, or of the distribution of defective items in a lot.) What if elements of Θ are vectors or even more complicated objects?

In any case, assume that the particular alternative hypothesis leads to rejection of the null hypothesis for large values of a *test statistic* T, say when $T > T_0$. How can one find the cut-off constant T_0? The traditional (but somewhat dated) approach is to specify a significance level α (a smallish probability, like 0.01 or 0.05) so that under H_0 the probability of its false rejection is α (or does not exceed α.) Then the probability of the event $T > T_0$ evaluated under the alternative represents the chance of the correct rejection and is called the *power* of the test.

A body of classical statistical literature deals with the problem of finding tests of a given significance level which have the largest power. To accomplish that typically a distribution from the alternative is to be fixed. An additional difficulty of the randomness hypothesis is that its amorphous alternative is enormous and cannot be fully described by a sensible finite-dimensional parameter set Θ_1.

A more modern approach (implemented in almost all statistical software packages) suggests that the main characteristic of a test rejecting the null hypothesis for large values of a statistic T, is the empirical significance level, i.e., the probability of the random variable T exceeding its observed value $T(obs)$ evaluated under the null hypothesis, $P(T > T_n(obs)|H_0)$. One of the immediate benefits of this concept, called the P-value, is that the classical test of level α obtains if the null hypothesis is rejected when the P-value is smaller that α.

The power function of a classical test can be expressed in terms of the P-value, but it involves calculation for $\theta \in \Theta_1$.

Since we believe that the P-value (not the significance level and not the power) is the most important practical characteristic of these procedures, each test in the suite results in such a value, and the collection of P-values from all the tests are consolidated into a vector reported to the consumer. If the distributions of two test statistics under the randomness hypothesis coincide, we consider them to be equivalent as the P-values are the same for both of these statistics.

Let n be the length of the string under testing. Each of tests in the suite is based on its statistic $T = T_n$ which, under the randomness assumption, has a desirably continuous distribution function $G(t) = G_n(t) = P(T \leq t)$ whose tail probabilities can be numerically evaluated. If a one-sided alternative corresponds to distributions of T which are stochastically larger than the distribution of T under the null hypothesis, then the P-value is $1 - G(T(obs))$. Its large values are indicative of the fact that the null hypothesis is false, i.e., they support the alternative hypothesis.

For example, in the classical goodness-of-fit test, T is the chi-squared statistic, and H_0 postulates the probabilities of a multinomial distribution as calculated from the randomness condition. As discussed later, the P-value can be obtained from the incomplete gamma-function, and its small values lead one to believe in the falsity of the null hypothesis. This type of statistic is common in the suite. On the other hand, statistics distributed as a mixture of chi-squared distributions with different degrees of freedom were deemed to be too inconvenient to work with.

For some tests the alternative to our randomness hypothesis may not necessarily be restricted to distributions of T which are stochastically larger (or smaller) than the distribution of T evaluated under this hypothesis. Then the two-sided alternatives can be more appropriate with the validity of the null hypothesis being in doubt for small values of $\min[G(T(obs)), 1 - G(T(obs))]$. An interpretation of P-values in this case is as a "degree of agreement" between the statistic and its "typical" value measured by the median $\hat{T}$ of its distribution (see Gibbons and Pratt, 1975, Rukhin, 2000). Section 2.2 gives an example.

When G is a discrete distribution and the alternative is one-sided, the P-value with the continuity correction is defined as $\frac{1}{2}P(T = T(obs)) + P(T > T(obs))$. Under the randomness hypothesis, these P-values have an approximate uniform distribution on the interval $(0, 1)$ (exactly uniform in the continuous case.) This can be tested, for example, by the classical Kolmogorov-Smirnov test.

To achieve P-values with a uniform distribution on the interval $(0, 1)$, when a discrete-valued statistic is used, the original string of length $n = NM$ is partitioned into N substrings each of length M. For each of these substrings the frequencies, $\nu_0, \nu_1, \ldots, \nu_K$, of values of the corresponding statistic within each of $K + 1$ chosen classes, $\nu_0 + \nu_1 + \ldots + \nu_K = N$, are evaluated. The theoretical probabilities $\pi_0, \pi_1, \ldots, \pi_K$ of these classes are determined from the (distribution of the test statistic.

The frequencies are aggregated by the χ^2-statistic,

$$\chi^2 = \sum_0^K \frac{(\nu_i - N\pi_i)^2}{N\pi_i},$$

which under the randomness hypothesis has an approximate χ^2-distribution with K degrees of freedom. The reported P-value cab be written as the incomplete gamma function.

This example (in which the exact distribution of T is a complicated sum of multinomial probabilities) demonstrates another difficulty of our testing problem. The exact distribution of T is usually difficult to find or it is too involved from the practical point of view. However the approximate (limiting as $n \to \infty$) distribution may be (more) tractable and available. By replacing G_n by this distribution one obtains approximate P-values. For insufficiently large n, these may lack accuracy when compared to the exact probabilities.

2 What have we learned: statistical tests which work well and some which do not

This section illustrates the concepts discussed in section 1 by using some basic tests in the suite.

2.1 Tests based on the properties of a random walk

Denote by $\epsilon_k, k = 1, 2, \ldots, n$ the underlying series of bits taking values 0 and 1 which is to be tested for randomness. In many situations it is more convenient to deal with the sequence $X_k = 2\epsilon_k - 1, k = 1, 2, \ldots, n$, with X_k taking values $+1$ or -1.

Quite a few statistical tests can be derived from the well-known limit theorems for the random walk, $S_n = X_1 + \cdots + X_n$. Under the randomness hypothesis, $(S_n + n)/2$ has the binomial distribution with parameters n and $p = 1/2$, which is not convenient to use when say, $n \geq 200$. However the classical Central Limit Theorem, according to which

$$\lim_{n\to\infty} P\left(\frac{S_n}{\sqrt{n}} \leq z\right) = \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2}\, du,$$

provides a useful approximation, which forms the foundation for the most basic monobit test of the null hypothesis that in a sequence of independent random variables X's or ϵ's the probability of ones is $1/2$.

More tests of randomness can be derived on the distribution of the maximum of the absolute values of the partial sums, $\max_{1\leq k\leq n} |S_k|$, and from the distribution of the number of visits within an excursion of the random walk S_k to a certain state. However, not all tests based on the probabilistic properties of random walk are equally suitable for randomness testing. For example, the limiting distribution of the proportion of time U_n that the sums S_k are non-negative, leads to a fairly weak test.

2.2 Discrete Fourier transform (spectral) test

The spectral test which appeared in the suite turned out to be troublesome. As it happened, it was not properly investigated, which resulted in a wrong statistic and a faulty constant. This fact was duly noticed by the cryptographic community (Kim, Umeno and Hasegawa, 2003, Killman, Schüth, Thumser and Uludag, 2004, Hamano, 2005). Now the original version is replaced by the following modification.

Let $X_k = \pm 1, k = 1, \ldots, M$, be a sequence of random bits. Denote

$$f_j = \sum_{k=1}^{M} X_k \exp\left\{ \frac{2\pi(k-1)j\mathbf{i}}{M} \right\},$$

$j = 0, \ldots, M/2 - 1$. Then $Ef_j = 0, Ef_j \bar{f}_{j'} = \delta_{jj'} M$. Here $\bar{z} = a - b\mathbf{i}$ is the complex conjugate of a complex number $z = a + b\mathbf{i}$. For a *fixed* m (i.e., m which does not depend on M) the joint distribution of the complex vectors $M^{-1/2}(f_1, \ldots, f_m)$ for large M is approximately the multivariate complex normal distribution. It follows that under the randomness hypothesis, $W = 2\sum_{k=1}^{m} mod_k^2/M, mod_k^2 = f_k \bar{f}_k$, has an approximate χ^2-distribution with $2m$ degrees of freedom.

These facts lead to the following procedure. Partition a string of length n, such that $n = MN$ into N substrings, each of length M. For each substring evaluate $mod_k^2, k = 1, \ldots, m$ (as in the original version of this test but for a much smaller m than $M/2 - 1$). For each $j = 1, \ldots, N$, calculate the statistic $W = W_j = 2\sum_{k=1}^{m} mod_k^2/M$. Reject the randomness hypothesis if the χ^2 goodness-of-fit test does not accept the χ^2-distribution with $2m$ degrees of freedom.

More exactly, choose a number $K + 1$ of disjoint intervals (classes), evaluate the theoretical probabilities $\pi_i, i = 0, 1, \ldots, K$ of the intervals according to this distribution, and form the familiar χ^2-statistic, $\chi^2 = \sum_{i=0}^{K} (\nu_i - N\pi_i)^2/(N\pi_i)$ with ν_i denoting empirical frequencies of W-values in i-th interval, $\nu_0 + \cdots + \nu_K = N$. The P-value can be given through the incomplete gamma-function,

An alternative way to calculate the P-value is to follow a suggestion in section 1.2. Namely, determine the median, $\hat{W}$, of the χ^2-distribution with $2m$ degrees of freedom, $Q(m, 0.5\hat{W}) = 0.5$. Then $0.5\hat{W} \approx m - 1/3$. The P-value corresponding to j-th substring is

$$\text{P-value } = \begin{cases} Q(m, 0.5W_j) + 1 - Q(m, \hat{W} - 0.5W_j) & W_j \geq \hat{W}, \\ 1 - Q(m, 0.5W_j) + Q(m, \hat{W} - 0.5W_j) & W_j < \hat{W}. \end{cases}$$

2.3 Non-overlapping and overlapping template matchings

Most conventional pseudo random number generators, such as the linear congruential generators and lagged-Fibonnaci generators used in IMSL, C++, and other packages, tend to show patterning due to their deterministic recursive algorithms. Because of this patterning, it is natural to investigate statistical tests based on the occurrences of words (patterns or templates).

We start here with the tests which utilize the observed numbers of words or the frequency of a given word in a sequence of length M. Let $\imath = (i_1, \ldots, i_m)$ be a given word (template or pattern, i.e. a fixed sequence of zeros and ones) of length m.

An important role belongs to the set $\{j, 1 \leq j \leq m, i_{j+k} = i_k, k = 1, \ldots, m-j\}$, which is the set of periods of $\imath$. For example, when $\imath$ corresponds to a run of m ones, $\{1, \ldots, m-1\}$ is the set of all periods. For *aperiodic* words $\imath$, this set is empty. Such words cannot be written as $\ell\ell \ldots \ell\ell'$ for a pattern ℓ shorter than $\imath$ with ℓ' denoting a prefix of ℓ. In this situation occurrences of $\imath$ in the string are necessarily non-overlapping.

Denote by $W = W(m, M)$ the number of occurrences of the given aperiodic pattern. If $(M - m + 1)/2^m = \lambda$, then $EW = (M - m + 1)2^{-m} = \lambda$. When both M and m tend to infinity, W has a Poisson limiting distribution with the parameter λ, i.e., $P(W = k) \to e^{-\lambda}\lambda^k/k!$ $k = 0, 1, \ldots$ (Barbour, Holst and Janson, 1992.) When the length m is fixed, the limiting distribution of standardized statistic W is normal. Each of these facts can be used for randomness testing.

This statistic is defined also for periodic patterns, but the accuracy of Poisson approximation is good only when $\imath$ does not have small periods. A test of randomness can be based on the number of possibly overlapping occurrences of templates in the string. If $U = U(m, M)$ is this number for a periodic word of length m then the asymptotic distribution of U is the compound Poisson distribution(the so-called Polya-Aeppli law). The probabilities of this distribution can be expressed in terms of confluent hypergeometric function. So to implement the test of randomness based on overlapping patterns, partition the string into N substrings and evaluate the empirical frequencies of occurrences of aperiodic or periodic patterns within each substring of length M comparing them to the theoretical probabilities via the χ^2-statistic.

Notice that M must be sufficiently large for validity of this test. For example, when $M = 1032$ and $\imath$ is a run of $m = 9$ ones, so that $\lambda = 1.9980468750$, the comparison of Polya-Aeppli probabilities and the exact probabilities (due to K. Hamano) is given in the Table 1.

Table 1. The exact probabilities and the Polya-Aeppli law probabilities when $M = 1032$ and $m = 9$

	exact probabilities	Pòlya-Aeppli probabilities
$P(U=0)$	0.367879	0.364091
$P(U=1)$	0.183940	0.185659
$P(U=2)$	0.137955	0.139381
$P(U=3)$	0.099634	0.100571
$P(U=4)$	0.069935	0.070431
$P(U=5)$	0.140657	0.139865

.

3 What we do not know yet: tests based on patterns, periodic or not

For a given set of words (patterns), it is of interest to determine the probability of the prescribed number of (overlapping) occurrences of these patterns in the text. This problem appears in different areas of information theory such as source coding and code synchronization. It is also important in molecular biology, in DNA analysis, and for gene recognition.

It is convenient now to consider a random text formed by realizations of letters chosen from a finite (not necessarily binary) alphabet. This setting can be used for a binary sequence if its substrings of a given length p represent the new letters, so that there are $q = 2^p$ such letters. Then the length n of q-nary sequence is related to the length n' of the original binary string by the formula, $n = n'/p$. This extension opens the possibility of choosing q in an optimal way.

To find words with prescribed frequencies one can use asymptotically normal estimates of word probabilities or the exact distributions obtained from generating functions (see, for example, Szpankowski, 2001). These results suggest that the probability for a given word $\imath$ to appear exactly r times in the string of length n can be approximated by the Poisson probability of the value r, when the Poisson parameter is $nP(\imath)$. Thus, the distribution of the number of words with given r can be expected approximately equal to that of the sum of Bernoulli random variables whose success probability is this Poisson probability. However, the more detailed structure of this distribution, in particular, the covariance of several such random variables, needed in the study of large sample efficiency, is less intuitive.

The approximate Poisson distribution for the number of missing words ($r = 0$) is alluded to in Marsaglia and Zaman (1993). It forms the basis of the so-called OPSO test of randomness in the Diehard Battery (Marsaglia 1996). This test takes non-overlapping substrings formed by zeros and ones of length $p = 10$ to represent the letters of the new alphabet, so that there are $q = 2^{10}$ new letters, which in general is not the optimal choice. In the OPSO test one counts the number of two-letter patterns (the original substrings of length $2p = 20$) which never occur. We consider the case of arbitrary m in the next section.

3.1 Tests of randomness based on the number of missing words

Let $\epsilon_1, \ldots, \epsilon_n$ denote a sequence of independent discrete random variables each taking values in the finite set $\mathbb{Q}$, say, $\mathbb{Q} = \{1, \ldots, q\}$, $P(\epsilon_i = k) = p_k, k = 1, \ldots, q$. Thus, the probability of the word $\imath = (i_1 \ldots i_m)$ is $P(\imath) = p_{i_1} \cdots p_{i_m}$. The situation when $p_k \equiv q^{-1}$ corresponds to the randomness hypothesis.

To determine efficient tests for randomness, we look at the alternative distributions of the alphabet letters which are close to the uniform in the sense that $p_k = q^{-1} + q^{-3/2}\eta_k$, $k = 1, \ldots, q$, $\sum_{k=1}^{q} \eta_k = 0$. We assume that as $n \to \infty$, $q \to \infty$ so that $n/q^m \to \lambda$ and for a positive κ, $\sum_k \eta_k^2/q \to \kappa$. Then $nP(\imath) \to \lambda$.

The first object of interest is the probability $\pi_\imath(n)$ that a fixed pattern $\imath$ is missing in the string of length n. To find this probability one can use the correlation polynomial of two patterns which was introduced by Guibas and Odlyzko (1981). It plays an important role in the study of the distribution of their frequencies.

For a complex z, let $F_\imath(z) = \sum_n \pi_\imath(n) z^{-n}$ be the probability generating function. Then in can be expressed as a ratio of two polynomials closely related to the correlation polynomials, and for any word $\imath$ the probability $\pi_\imath(n)$ can be found by comparing the coefficients in the series expansions of $F_\imath(z)$.

For example when $m = 2$, $\jmath = (i,k), i \neq k$, then $F_\jmath(z) = z^2/(p_i p_k + (z-1)z)$. With $s = 1/2 + \sqrt{1/4 - p_i p_k}$, $t = 1/2 - \sqrt{1/4 - p_i p_k}$, $\pi_\jmath = (s^{n+1} - t^{n+1})/(s-t)$. These formulas lead to very accurate answers for the expected value and the variance. For example when $n = 2^{21}$, $q = 2^{10}$ (so that $\lambda = 2$),

$$\pi_{(i,i)} = 0.13559935200020, \; \pi_{(i,k)} = 0.13533502510527.$$

The asymptotic approximation for the probabilities $\pi_\jmath^t(n)$ for a word $\jmath$ of length m and of period t is $\pi_\jmath^t(n) \approx e^{-\lambda}(1 + \lambda q^{-t})$. For aperiodic words, $\pi_\jmath^\infty(n) \sim e^{-\lambda}(1 - (2m-1)\lambda/(2q^m) + (m-1)/q^m)$. To get the formula for the expected value of the number of missing m-words, X^0 let $N_t = N_t(m), t = 1, \ldots, m-1, \infty$, denote the total number of words of the period t ($t = \infty$ corresponds to aperiodic words). Then $\sum N_t = N^m$, and as $q \to \infty$ for $t = 1, \ldots, m-1$, $N_t \sim q^t, N_\infty \sim q^m$. One gets

$$EX^0 = \sum_t N_t \pi_\jmath^t(n) \approx e^{-\lambda} q^m + e^{-\lambda} \left[m - 1 - \frac{\lambda}{2} \right].$$

In the example when $m = 2, n = 2^{21}$, $q = 2^{10}$, the exact value of the mean is $EX^0 = 141909.3299555$, and the approximate formula gives 141909.3299551. The formula for the variance can be obtained from the probabilities $\pi_{\imath\jmath}^{00}(n) = P\,$(words $\imath$ and $\jmath$ are missing) which can be found from the probability generating function technique. See Rukhin (2001, 2002) for details.

After the number X^0 of missing two letter words in the string of length n has been determined, one evaluates the ratio, $(X^0 - EX^0)/\sqrt{Var(X^0)}$, which leads to the P-value obtained from the standard normal distribution.

3.2 Testing randomness via words with a given frequency

More powerful tests can be derived by using the observed numbers of words which appear in a random text a prescribed number of times (i.e. which are missing, appear exactly once, exactly twice, etc.) In practice these statistics are easier to evaluate than the empirical distribution of occurrences of all m-words.

It turns out that such tests can be obtained by techniques of the previous section which lead to the formula for the expected value of the number of m-words, which occur exactly r times in a random string of length n, $X^r = X_n^r$, A surprising fact is that the asymptotic behavior of the expected value and of the covariance matrix is the same for overlapping and non-overlapping occurrences, i.e., when the word occurrences are counted in the non-overlapping m-blocks. Therefore, the form of the following optimal test coincides with that in Kolchin, Sevastyanov and Chistyakov (1978) who give the formulas for the first two moments of the joint distribution of the number of words appearing a prescribed number of times when the frequencies of these words are independent.

To derive the optimal test of the null hypothesis $H_0 : \eta_i \equiv 0$, we look at the class of linear test statistics of the form $S = \sum_{r=0}^{R} w_r(X^r - EX^r)$ for a fixed positive integer R. The (Pitman) efficiency of this statistic can be obtained from the fact that S is asymptotically normal both under the null hypothesis and the alternative $H_1 : \kappa > 0$. This efficiency is determined by the normalized distance between the means under the null hypothesis and under the alternative, divided by the standard deviation (which is common to the null hypothesis and the alternative). This test is asymptotically optimal not only within the class of linear statistics, but in the class of all functions of $X^0, \ldots, X^R$.

The following table for $R = 0, \ldots, 9$, gives the value of $\lambda = \lambda^\star$, which maximizes the efficiency and the corresponding optimal weights $\tilde{w} = w/\sum w_r$ normalized so that their sum is equal to one.

Table 2. Optimal weights and the optimal $\lambda^\star$

R	$\lambda^\star$	$\tilde{w}$
0	3.59	1
1	4.77	$[0.62, 0.38]$
2	5.89	$[0.47, 0.33, 0.20]$
3	6.98	$[0.39, 0.29, 0.20, 0.14]$
4	8.06	$[0.33, 0.25, 0.19, 0.14, 0.09]$
5	9.13	$[0.29, 0.23, 0.18, 0.14, 0.09, 0.07]$
6	10.17	$[0.25, 0.21, 0.18, 0.14, 0.09, 0.07, 0.06]$
7	11.21	$[0.23, 0.19, 0.16, 0.14, 0.09, 0.07, 0.06, 0.05]$
8	12.24	$[0.21, 0.18, 0.16, 0.14, 0.09, 0.07, 0.06, 0.05, 0.04]$
9	13.26	$[0.19, 0.17, 0.16, 0.14, 0.09, 0.07, 0.06, 0.05, 0.04, 0.03]$

To implement this test on the basis of a string of binary bits for a fixed R, as discussed in the beginning of this section, choose a positive integer p, such that $n \approx 2^{mp}\lambda^\star$, and take all substrings of length p formed by zeros and ones to represent the letters of the new alphabet of the size $q = 2^p$. The numbers X^r of m-letter patterns (the original non-overlapping consecutive substrings of length $2m$), which occurred r times with the weights from the table lead to the asymptotically optimal test. In particular, the most efficient test based on the number of missing words $m = 2, R = 0$ arises when $\lambda^\star = 3.594..$, which means that the best relationship between q and n, is $n \approx 3.6q^2$ Extensions of these results to Markov dependent alternatives are in Rukhin (2006).

4 What are the most important open problems: data compression and randomness testing

It is desirable to develop tests based on patterns suggested by the data themselves. A powerful heuristic idea is that random sequences are those that cannot be compressed or those that are most complex. However its practical implementation is limited by scarcity of relevant compression code based statistics whose (approximate) distributions can be evaluated.

4.1 Linear complexity for testing randomness

Here we look at linear complexity which is related to one of the main components of many keystream generators, namely, Linear Feedback Shift Registers (LFSR). Such a register consists of L delay elements each having one input and one output. If the initial state of LFSR is $(\epsilon_{L-1}, \ldots, \epsilon_1, \epsilon_0)$, then the output sequence, $(\epsilon_L, \epsilon_{L+1}, \ldots)$, satisfies the following recurrent formula for $j \geq L$

$$\epsilon_j = (c_1\epsilon_{j-1} + c_2\epsilon_{j-2} + \cdots + c_L\epsilon_{j-L}) \mod 2.$$

Here $c_1, \ldots, c_L$ are coefficients of the so-called connection polynomial corresponding to a given LFSR. The *linear complexity* $\mathcal{L} = \mathcal{L}_n$ of a given sequence $\epsilon_1, \ldots, \epsilon_n$, is defined as the length of the shortest LFSR that generates it as first n terms. The possibility of using the linear complexity characteristic for testing randomness is based on the Berlekamp-Massey algorithm, which provides an efficient way to evaluate the connection polynomial for finite strings.

When the binary n-sequence is truly random, the formulas for the mean, $\mu_n = E\mathcal{L}_n$, and the variance are known. However, the asymptotic distribution as such does not even exist; one has to treat the cases, n even, and n odd, separately with two different limiting distributions. Both of these distributions can be conjoined in a discrete distribution obtained via a mixture of two geometric random variables (one of them taking only negative values).

The monograph of Rueppel (1986) gives the distribution of the random variable $\mathcal{L}_n$, the linear complexity of a random binary string, which can be used to show that

$$T_n = (-1)^n[\mathcal{L}_n - \xi_n] + \frac{2}{9},\ \xi_n = \frac{n}{2} + \frac{4 + r_n}{18},$$

can be used for testing randomness. The sequence T_n converges in distribution to the random variable T whose distribution is skewed to the right. While $P(T = 0) = 1/2$, $P(T = k) = 2^{-2k}$, for $k = 1, 2, \ldots$, and $P(T = k) = 2^{-2|k|-1}, k = -1, -2, \ldots$, which provide easy formulas for the P-values.

In view of the discrete nature of this distribution one can use the strategy described in section 1.2 for a partitioned string. A more powerful test which efficiently uses the available data was suggested by Hamano, Sato and Yamamoto (2009). It is based on the statistic $\sum |j/2 - \mathcal{L}_j|$, which can be interpreted as the sum of areas of triangles formed by vertexes $(j, \mathcal{L}_j), (j+1, \mathcal{L}_{j+1}), j = 0, 1, 2, \cdots$, around the line $\mathcal{L}_j = j/2$.

4.2 Tests based on data compression

The original suite attempted to develop a randomness test based on the Lempel-Ziv algorithm (1977) of data compression via parsing of the text. Let W_n represent the number of words in the parsing of a binary random sequence of length n according to this algorithm. Aldous and Shields (1988) have shown that $EW_n/(n/\log_2 n) \to 1$, so that expected compression can be asymptotically approximated by $n/\log_2 n$. Moreover,

$$P\left(\frac{W_n - EW_n}{\sqrt{Var(W_n)}} \le w\right) \to \Phi(w).$$

The behavior of $Var(W_n)$ was elucidated by Kirschenhofer, Prodinger, and Szpankowski (1994) who proved that $Var(W_n) \sim (n[C+\delta(\log_2 n)])/\log_2^3 n$, where $C = 0.26600$ (to five significant places) and $\delta(\cdot)$ is a slowly varying continuous function with mean zero, $|\delta(\cdot)| < 10^{-6}$.

One of the tests in the original version of the suite compressed the sequence using the Lempel-Ziv algorithm. If the reduction is statistically significant when compared to the expected result, one can declare the sequence to be nonrandom. It was expected that the ratio $(W - n/\log_2 n)/\sqrt{0.266n/\log_2^3 n}$, where W is the number of words obtained, would provide the P-value corresponding to the two-sided alternative. Unfortunately, this test failed because the normal approximation was too poor, i.e., the asymptotic formulas are not accurate enough for values of n of the magnitude encountered in testing random number generators.

More practical turned out to be the so-called "universal" test introduced by Maurer (1992). The test requires a long (of the order $10 \cdot 2^L + 1000 \cdot 2^L$ with $6 \le L \le 16$) sequence of bits which it divides into two stretches of L-bit blocks: D, $D \ge 10 \cdot 2^L$, initialization blocks and K, $K \approx 1000 \cdot 2^L$ test blocks. The test looks back through the entire sequence while inspecting the test segment of L-bit blocks, checking for the nearest previous exact match and recording the distance (in number of blocks) to that previous match. The algorithm computes the logarithm of all such distances for all the L-bit templates in the test segment (giving effectively the number of digits in the binary expansion of each distance), and averages over all the expansion lengths by the number of test blocks K to get the test statistic F_n. A P-value is obtained from the normal error function based on the standardized version of the statistic, with the test statistic's mean EF_n equal to that of $\log_2 G$, where G is a geometric random variable with the parameter $1 - 2^{-L}$.

The difficult part is determination of the variance $Var(F_n)$. There are several versions of empirical approximate formulas, $Var(F_n) = c(L,K)Var(\log_2 G)/K$, where $c(L,K)$ represents the factor which takes into account dependent nature of the occurrences of templates, The latest of the approximations belonging to Coron and Naccache (1998) has the form $c(L,K) = 0.7 - 0.8/L + (1.6 + 12.8/L)K^{-/L}$. However, these authors report that "the inaccuracy due to [this approximation] can make the test to be 2.67 times more permissive than what is theoretically admitted."

The prospects for better approximations, in particular for the exact variances $Var(W_n)$ or $Var(F_n)$ do not look very good. In view of this fact, it may be advisable to test the randomness hypothesis by verifying normality of the observed values F_n assuming that the variance is unknown. This can be done via a classical statistical technique, namely the t-test. The original sequence must be partitioned in a number N (say $N \leq 20$) of substrings on each of which the value of the universal test statistic is evaluated (for the same value of parameters K, L and D). The sample variance is calculated, and the P-value is obtained from the t-distribution with $N-1$ degrees of freedom.

The most interesting randomness test would be based on Kolmogorov's definition of complexity which is the length of the shortest (binary) computer program that describes the string. One of the universal Turing machines is supposed to represent the computer which could use this description to exhibit this string after a finite amount of computation. As was argued, the random sequences are the most complex ones, so if Kolmogorov's complexity were computable, a randomness test would reject the null hypothesis for its small values. Unfortunately, this complexity characteristic is not computable (Cover and Thomas, 1991), and there is no hope for a test which is directly based on it.

Pursuing the idea of using data compression codes as randomness testing statistics, let $\mathbb{Q}$ be a finite alphabet of size q. A data compression method consists of a collection of mappings ϕ_n of $\mathbb{Q}^n, n = 1, 2, \ldots$ into a set of all finite binary sequences, such that for $\imath, \jmath \in \mathbb{Q}^n$, $\imath \neq \jmath$, one has $\phi_n(\imath) \neq \phi_n(\jmath)$. This means that a message of any length n can be both compressed and decoded.

For a given compression code, the randomness hypothesis is accepted on the basis of the string $\epsilon_1, \ldots, \epsilon_n$ if the length T_n of $\phi_n(\epsilon_1, \ldots, \epsilon_n)$ is large enough. Ryabko and Monarev (2005) show that the choice of the cut-off constant $T_0 = n \log q + \log \alpha - 1$ leads to a test of significance level α. A code is universal if for any ergodic stationary source Γ, the ratio T_n/n converges with probability one to the entropy of Γ. This entropy equals to $\log q$ if the randomness hypothesis is true, and it is smaller than $\log q$ under any alternative that can be modeled by an ergodic stationary process. For universal codes the power of the corresponding test tends to one as n increases. However, finding non-trivial compression codes with known distributions of the codewords length (so that P-value can be evaluated) is quite difficult.

5 What are the prospects for progress: Concluding remarks

To sum up, there are major challenges in the area of empirical randomness testing. It may be a bit surprising that so many available procedures are based on the one hundred years old χ^2-test. Since this area is so important, one can expect more stringent methods based on new ideas. In particular, a study of overlapping spatial patterns is of great interest, as it may lead to such procedures.

6 References

1. Aldous, D. and Shields, P. A diffusion limit for a class of randomly-growing binary trees. *Probability Theory and Related Fields* 79, 1988, 509-542.

2. Barbour A. D., Holst, L. and Janson, S. *Poisson Approximation*, Clarendon Press, Oxford, 1992.

3. Coron, J-S. and Naccache, D. An accurate evaluation of Maurer's universal test. Proceedings of SAC'98, Lecture Notes in Computer Science, Springer, Berlin, 1998.

4. Cover, T. and Thomas, J. Elements of Information Theory, J. Wiley, New York, NY, 1991,

5. Gibbons, J. and Pratt, J. P-values: interpretations and methodology. *American Statistician* 29, 1975, 20–25.

6. Guibas, L. J. and Odlyzko, A. M. Strings overlaps, pattern matching and nontransitive games. *J. Comb. Theory* A 30, 1970, 183–208.

7. Gustafson, H., Dawson, E., Nielsen, L., and Caelli, W. A computer package for measuring the strength of encryption algorithms. *Computers and Security* 13, 1994, 687-697.

8. Hamano, K. The distribution of the spectrum for the discrete Fourier transform test included in SP800-22. *IEICE Trans. Fundamentals* E88, 2005, 67-73.

9. Hamano, K., Sato, F., and Yamamoto, H. A new randomness test based on linear complexity profile. *IEICE Trans. Fundamentals* E92, 2009, 166-172.

10. Killmann, W., Schüth, J., Thumser, W. and Uludag, I. A note concerning the DFT test in NIST special publication 800-22.T-Systems Integration, Technical Report, 2004.

11. Kim, S.-J., Umeno, K. and Hasegawa, A. Corrections of the NIST statistical suite for randomness, IEICE Technical Report, ISEC2003-87, 2003.

12. Kirschenhofer, P., Prodinger, H., and Szpankowski, W. Digital Search Trees Again Revisited: The Internal Path Length Perspective. *SIAM Journal on Computing* 23, 1994, 598-616.

13. Knuth, D. E. *The Art of Computer Programming*, Vol. 2, 3rd ed. Addison-Wesley Inc., Reading, MA. 1998.

14. V. F. Kolchin, B. A. Sevast'yanov, and V. P. Chistyakov. *Random Allocations.* Whinston Sons, Washington, DC, 1978.

15. L'Ecuyer, P. and Simard, R. TestU01: A C library for empirical testing of random number generators, *ACM Transactions of Mathematical Software.* 33, 4, Article 22, 2007.

16. Marsaglia, G. *Diehard: A battery of tests for randomness.*
    http://stat.fsu.edu/ geo/diehard.html 1996.

17. Marsaglia, G. and Zaman, A. Monkey tests for random number generators. *Computers&Mathematics with Applications* 9, 1993, 1–10.

18. Maurer, U. M. A universal statistical test for random bit generators. *Journal of Cryptology* 5, 1992, 89-105.

19. Murphy, S. The power of NIST's statistical testing of AES candidates. 2000. http://www.cs.rhbnc.ac.uk/ sean/StatsRev.ps

20. Rueppel, R. *Analysis and Design of Stream Ciphers.* Springer, Berlin, 1986.

21. Rukhin, A. L. Admissibility: survey of a concept in progress. *International Statistical Review* 63, 1995, 95–115.

22. Rukhin, A. L. Testing randomness: a suite of statistical procedures. *Theory of Probability and its Applications* 45, 2000, 137–162.

23. Rukhin, A. L. Pattern correlation matrices and their properties. *Linear Algebra and its Applications* 327, 2001, 105–114.

24. Rukhin, A. L. Distribution of the number of words with a prescribed frequency and tests of randomness. *Advances in Applied Probability* 34, 2002, 775–797.

25. Rukhin, A. L. Pattern correlation matrices for Markov sequences and tests of randomness. *Theory of Probability and its Applications* 51, 2006, 712-731.

26. Rukhin, A. L., Soto, J., Nechvatal, J., Smid, M., Levenson, M., Banks, D., Vangel, M., Leigh, S., Vo, S., Dray, J. A statistical test suite for the validation of cryptographic random number generators. Special NIST Publication, NIST, Gaithersburg, 2000.

27. Ryabko, B. Ya. and Monarev, V. A. Using information theory approach to randomness testing. *Journal of Statistical Planning and Inference* 133, 2005, 95–110.

28. Soto, J. and Bassham, L. Randomness testing of the advanced encryption standard finalist candidates. *Proceedings of AES Conference*, 2001, http://csrc.nist.gov/ publications/nistir/ir6483.pdf

29. Szpankowski, W. *Average Case Analysis of Algorithms on Sequences.* Wiley-Interscience, New York, 2001.

30. Ziv, J. and Lempel, A. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23, 1997, 337-343.