TESTING RANDOMNESS: A SUITE OF STATISTICAL PROCEDURES

Andrew L. Rukhin
Department of Mathematics and Statistics
UMBC
Baltimore, MD 21250 USA

A suite of procedures designed to test randomness of binary sequences is discussed.

Keywords Correlation polynomial, Entropy, Distribution of m-patterns, P-values, χ^2-distribution.

1 Introduction

The problem of randomness testing discussed in this paper is motivated by attempts to assess the quality of different random number generators. The wide use of public key cryptography has led to the interest in testing for randomness a binary string produced by such generators. The evaluation of random nature of outputs produced by various generators became vital for communications industry where digital signatures and key management are crucial for information processing and for computer security.

A number of classic tests of randomness is reviewed in Knuth (1998). However, most of these tests may pass patently nonrandom sequences (see discussion in Marsaglia, 1985). The most popular battery of tests for randomness, the Diehard Battery, (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.

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 battery of modern stringent procedures. The resulting suite (Rukhin et al, 2000) was successfully applied to pseudo-random binary sequences produced by current generators, which are used in encryption as well as in scientific computing. The aim at NIST was not to identify the best generator, but rather to provide a user with a characteristic allowing to make an informed choice about the generator. The key selection criteria for inclusion in the suite were that the test states its result as a P-value, that the mathematics behind the test be applicable in the finite-sequence domain (possibly through simulation), and that the test not essentially duplicate a test that is already in the suite.

A list of some core tests designed for this purpose follows. Most of them are based on known results of probability theory and information theory; only few of these procedures are new. Almost all of the tests are applicable for a wide range of binary strings size n and, thus, exhibit greater flexibility.

2 P-values: one-sided alternatives versus two-sided alternatives

From the point of view of statistical hypotheses testing the principal difficulty of testing randomness is that this null hypothesis is typically false. Indeed, this is certainly the case with all pseudorandom number generators which are based on recursive formulas. In view of this fact, one may expect only a measure of "degree of non-randomness" attested to by a given string.

This measure for each test in the suite is a P-value, (empirical significance level), and the collection of P-values from all the tests forms the characteristic reported to the consumer. All considered tests are based on a statistic $T = T_n$ which under randomness assumption has a non-degenerate, desirably continuous, limiting distribution function G whose tail probabilities can be numerically evaluated. The number $1 - G(T_n(obs)) = P(T \geq T_n(obs))$ provides then the approximate P-value of the null hypothesis of randomness against the one-sided alternative corresponding to distributions of T being stochastically larger than the distribution of T evaluated under the null hypothesis. For example, in the classical example, when T is the chi-squared statistic, the P-value, $P(T_n \geq T_n(obs))$, can be obtained as the incomplete gamma-function, and its small values lead one to believe in the falsity of the null hypothesis. On the other hand, statistics with the distribution being a mixture of the chi-squared distributions with different degrees of freedom, were deemed to be too inconvenient to work with.

Since 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. one may look at two-sided alternatives. Then the validity of the null hypothesis is in doubt for small values of $\min[G(T_n(obs)), 1 - G(T_n(obs))]$. We suggest to interpret P-values in this case as a "degree of agreement" between the statistic and its "typical" value, which is to be measured by the median $\hat{T}$ of its distribution (see Gibbons and Pratt, 1975). More precisely, the P-value is defined as $P(\left|T_n - \hat{T}\right| \geq \left|T_n(obs) - \hat{T}\right|)$. Thus

$$P\text{-value} = \begin{array}{ll} 1 - G(T_n(obs)) + G(2\hat{T} - T_n(obs)), & T_n(obs) \geq \hat{T}, \\ 1 + G(T_n(obs)) - G(2\hat{T} - T_n(obs)), & T_n(obs) < \hat{T}. \end{array}$$

Observe that as a function of $T_n(obs)$, the P-value is an increasing function for

$T_n(obs) < \hat{T}$, attains its maximum, 1, at $\hat{T}$ and decreases afterwards. It follows that if, say, T is positive and $\hat{T} > T(obs)$, then P-value cannot be smaller than $1 - G(2\hat{T})$. The two-sided test based on T is not appropriate in this situation.

When G is a discrete distribution with a one-sided alternative, the P-value is defined as $\frac{1}{2}P(T = T_n(obs)) + P(T > T_n(obs))$. Under the randomness hypothesis, these P-values have approximate uniform distribution on the interval $(0, 1)$. This distribution is exactly uniform in the continuous case, and this uniformity is tested by Kolmogorov-Smirnov test in the suite. To achieve P-values with the uniform distribution on the interval $(0, 1)$, when a discrete-valued statistic is used, the original string 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 (discrete) distribution of the test statistic.

The frequencies are conjoined 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 the approximate χ^2-distribution with K degrees of freedom. The reported P-value is

$$\frac{\int_{\chi^2(obs)}^{\infty} e^{-u/2} u^{K/2-1}\, du}{\Gamma(K/2)\, 2^{K/2}} = 1 - \mathbf{P}\left(\frac{K}{2}, \frac{\chi^2(obs)}{2}\right),$$

with $\mathbf{P}(a, x)$ denoting the incomplete gamma function

$$\mathbf{P}(a, x) = \frac{\int_0^x e^{-u} u^{a-1}\, du}{\Gamma(a)}.$$

3 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 some situations it is more convenient to deal with the sequence $X_k = 2\epsilon_k - 1, k = 1, 2, \ldots, n$, with X_k are taking values $+1$ or -1.

Many tests can be derived from the well-known limit theorems for the random walk, $S_n = X_1 + \cdots + X_n$. For example, the classical Central Limit Theorem, according to which

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

forms the foundation for the most basic monobit test of the null hypothesis that in a sequence of i.i.d. random variables X's or ϵ's the probability of ones is $1/2$.

Other results based on the approximation of the distributions of functionals of the random walk by those of the Brownian motion also can be useful (see Skorokhod and Slobodenyuk, 1970). For example, the limiting distribution of the proportion of time U_n that the sums S_k are non-negative,

$$\lim_{n\to\infty} P\left(\frac{U_n}{n} < z\right) = \frac{2}{\pi} arcsin\sqrt{z},\ 0 < z < 1.$$

can be used for randomness testing, although is seems to lead to a weaker test than the following procedure.

3.1 A test based on the maximum of the absolute values of the random walk

The suggested test is based on the distribution of the maximum of the absolute values of the partial sums, $\max_{1\le k\le n}|S_k|$. Using Theorem 2.6, p 17 of Revesz (1990) one obtains

$$P\left(\max_{1\le k\le n}|S_k| \ge t\right) = 1 - \sum_{k=-\infty}^{\infty} P\left((4k-1)t < S_n < (4k+1)t)\right)$$

$$+ \sum_{k=-\infty}^{\infty} P\left((4k+1)t < S_n < (4k+3)t)\right).$$

As with probability one $|S_n| \le n$, the summation in the first sum above can be restricted to values k such that $4|k| < n/t+1$ and in the second sum $(-n/t-3) < 4k < n/t - 1$.

This is an *exact* formula and, as such, could be used for fairly short strings. Indeed $(S_n + n)/2$ has the binomial distribution with parameters n and $p = 1/2$, and all probabilities above can be written in terms of this distribution. However even for small values of n, the following approximation gives pretty good numerical results,

$$\lim_{n\to\infty} P\left(\max_{1\le k\le n}|S_k| \le \sqrt{n}z\right) = \frac{1}{\sqrt{2\pi}}\int_{-z}^{z}\sum_{k=-\infty}^{\infty}(-1)^k \exp\left\{-\frac{(u-2kz)^2}{2}\right\} du$$

$$= \frac{4}{\pi}\sum_{j=0}^{\infty}\frac{(-1)^j}{2j+1}\exp\left\{-\frac{(2j+1)^2\pi^2}{8z^2}\right\} = H(z),\quad z > 0. \tag{1}$$

With $z = \max_{1\le k\le n}|S_k|(obs)/\sqrt{n}$, the approximate P-value according to this formula is $1 - H(z)$, and the randomness hypothesis is rejected for large values of z. The series in the last formula in (1) converges fast and should be used for numerical calculation only for small values of z. The following formula for $H(z)$ is to be used for large values of z.

For a fixed positive z the function

$$g(u) = \frac{1}{\sqrt{2\pi}} \sum_{k=-\infty}^{\infty} (-1)^k \exp\left\{-\frac{(u-2kz)^2}{2}\right\}$$

is even and $4z$-periodic. More precisely, for any integer ℓ

$$g(u + 2\ell z) = (-1)^\ell g(u),$$

and the validity of the formulas in (1) can be derived by integration of the Fourier expansion of this function into $\cos \frac{\pi u}{2z}(2k+1), \quad k = 0, 1, \ldots$. Also

$$H(z) = \frac{1}{\sqrt{2\pi}} \int_{-z}^{z} \sum_{k=-\infty}^{\infty} (-1)^k \exp\{-\frac{(u-2kz)^2}{2}\}\, du$$

$$= 2\,[\Phi(z) - \Phi(-z)] + 2\sum_{k=0}^{\infty}(-1)^k\,[\Phi((2k+1)z) - \Phi((2k+1)z)]$$

$$\approx 1 - 4\,[\Phi(3z) - \Phi(5z)] \approx 1 - \frac{4}{3\sqrt{2\pi}z}\exp\{-\frac{9z^2}{2}\}, \quad z \to \infty.$$

Note that a dual test based on (1) can be derived form the "reversed time" random walk with $S'_k = X_n + \cdots + X_{n-k+1}$. With this definition, the interpretation of the tests results is to be modified by replacing " the early stages " by "the late stages".

3.2 A test based on the the number of visits of the random walk

Another test of randomness can be obtained from the distribution of the number of visits within an excursion of the random walk $S_k = X_1 + \cdots + X_k$ to a certain state.

Consider this random walk S_k as a sequence of excursions

$$(i, \ldots, \ell) \ : S_{i-1} = S_{\ell+1} = 0, \ S_k \neq 0 \text{ for } i \leq k \leq \ell.$$

We denote by J the total number of such excursions in the string. The limiting distribution for this (random) number J (i.e. the number of zeros among the sums $S_k, k = 1, 2, \ldots, n$ when $S_0 = 0$) is known

$$\lim_{n\to\infty} P\left(\frac{J}{\sqrt{n}} < z\right) = \sqrt{\frac{2}{\pi}} \int_0^z e^{-u^2/2}\, du, \quad z > 0.$$

One can reject the randomness hypothesis right away if J is too small, i.e. if the following P-value is small,

$$P\left(J < J(obs)\right) \approx \sqrt{\frac{2}{\pi}} \int_0^{J(obs)/\sqrt{n}} e^{-u^2/2}\, du = \mathbf{P}\left(\frac{1}{2}, \frac{J^2(obs)}{2n}\right).$$

Otherwise the number of visits of the random walk S to a certain state is evaluated.

Let $\xi(x)$ be the number of visits to $x, x \neq 0$, during one excursion. The following distribution of $\xi(x)$ can be found in Theorem 9.7, p 96 of the book of Revesz (1990) (see also Baron and Rukhin, 1999),

$$P\left(\xi(x)=0\right)=1-\frac{1}{2|x|} \tag{2}$$

and for $k=1,2,\ldots$

$$P\left(\xi(x)=k\right)=\frac{1}{4x^2}\left(1-\frac{1}{2|x|}\right)^{k-1}. \tag{3}$$

This distribution means that $\xi(x)=0$ with probability $1-1/2|x|$, otherwise (with probability $1/2|x|$) it equals to a geometric random variable with the parameter $1/2|x|$.

It is easy to see that $E\xi(x)=1$ and $Var(\xi(x))=4|x|-2$. A useful formula

$$P(\xi(x)\geq a+1)=2xP(\xi(x)=a+1)=\frac{1}{2|x|}\left(1-\frac{1}{2|x|}\right)^{a},\quad a=0,1,2,\ldots. \tag{4}$$

follows from (3).

This result can be used for randomness testing in the following way. For a "representative" collection of x-values (say, $1\leq x\leq 7$ or $-7\leq x\leq -1$), evaluate the observed frequencies $\nu_k(x)$ of the number k of visits to the state x during J excursions which occur in the string. Thus $\nu_k(x)=\sum_{j=1}^{J}\nu_k^j(x)$ with $\nu_k^j(x)=1$, if the number of visits to x during the jth excursion ($j=1,\ldots,J$) is exactly equal to k, and $\nu_k^j(x)=0$, otherwise. Pool the values of $\xi(x)$ into classes, say, $k=0,1,\ldots,4$ with an additional class $k\geq 5$. Theoretical probabilities of these classes can be obtained from (2), (3) and (4), namely,

$$\pi_0(x)=P\left(\xi(x)=0\right)=1-\frac{1}{2|x|};$$

$$\pi_k(x)=P\left(\xi(x)=k\right)=\frac{1}{4x^2}\left(1-\frac{1}{2|x|}\right)^{k-1}, k=1,\ldots,4;$$

$$\pi_5(x)=P(\xi(x)\geq 5)=\frac{1}{2|x|}\left(1-\frac{1}{2|x|}\right)^{5}.$$

Compare the frequencies $\nu_k(x)$ to the theoretical ones by the χ^2-test,

$$\chi^2(x)=\sum_{k=0}^{5}\frac{(\nu_k(x)-J\pi_k(x))^2}{J\pi_k(x)},$$

which for any x under the randomness hypothesis must have approximately the χ^2-distribution with 5 degrees of freedom. This is a valid test when $J\min\pi_k(x)\geq 5$, i.e. if $J\geq 1,000$. If this condition does not hold, one has to pool values of $\xi(x)$ into larger classes.

4 Runs Tests of Randomness

The classical definition of a *run*, say, of zeros is a succession of one or more zeros which are followed and preceded either by one or by no symbol at all. It is worth noticing that this is not the only possible definition of a run. Sometimes a run is not allowed to start or to end with no symbol at all. There is also a different definition of the *length* of a run, or of a run of a given length. For example, Feller defines a run of length r as a recurrent event (see Feller, 1968, Vol 1, Third ed., Sec XIII.1 and Sec XIII.7.) Namely, such a run is a non-overlapping with other runs. The advantage of Feller's definition is that the moments of the occurrences of a run of length m admit a fairly simple generating function (see Sec 7 of Chapter XIII in Feller, 1968) with explicit formulas for the mean $\mu = 2^{m+1} - 2$ and the variance $\sigma^2 = 2^{2(m+1)} - (2m+1)2^{m+1} - 2$.

For a fixed m the number N_m of Feller defined runs of length m is approximately normal,

$$\lim_{n\to\infty} P\left(\frac{(N_m - n/\mu)\mu^{3/2}}{\sqrt{n}\sigma} < z\right) = \Phi(z). \tag{5}$$

Therefore with $z(obs) = \sqrt{\mu}(\mu N_m(obs) - n)/(\sigma\sqrt{n})$ the P-value is $2(1 - \Phi(|z(obs|))$.

A similar limit theorem holds for the classical definition of a run of a fixed length m. If M_m denotes the total number of runs of length m in a string of size n with n_1 ones and $n_0 = n - n_1$ zeros, then

$$EM_m = \frac{n_0(n_0+1)n_1(n_1-1)\cdots(n_1-m+1)}{n(n-1)\cdots(n-m)}$$

and

$$Var(M_m) = \frac{n_0^2(n_0^2-1)n_1(n_1-1)\cdots(n_1-2m+1)}{n(n-1)\cdots(n-2m-1)}$$

$$+\frac{n_0(n_0+1)n_1(n_1-1)\cdots(n_1-m+1)}{n(n-1)\cdots(n-m)}\left[1-\frac{n_0(n_0+1)n_1(n_1-1)\cdots(n_1-m+1)}{n(n-1)\cdots(n-m)}\right]$$

with $(M_m - EM_m)/\sqrt{Var(M_m)}$ being approximately standard normal.

The situation changes when the length m is allowed to grow as n increases. Let $W = W(m,n)$ be the number of Feller defined, non-overlapping runs of length m. Then if $n, m \to \infty$ so that

$$\frac{n}{2^m} \to \lambda > 0, \tag{6}$$

then W has a Poisson limit

$$P(W = k) \to e^{-\lambda}\frac{\lambda^k}{k!} \quad k = 0, 1, \ldots. \tag{7}$$

See Barbour, Holst and Janson (1992, Sec 8.4).

However, the number $\tilde{W} = \tilde{W}(m,n)$ of overlapping runs of length m in the string has a different limiting distribution, namely a compound Poisson distribution (the so-called Pòlya-Aeppli distribution), with moment generating function

$$E \exp\{t\tilde{W}\} \to \exp\left\{\frac{\lambda(e^t-1)}{1-e^t/2}\right\} \tag{8}$$

when (6) holds. To use these results one adopts the strategy discussed in Section 2, namely, the string is partitioned into N substrings and the empirical frequencies within each substring are conjoined by the χ^2-statistic.

The limiting distribution of the total number V_n of runs (in the classical setting) for the fixed proportion of ones $\lambda = n_1/n$ is known to be normal

$$\lim_{n\to\infty} P\left(\frac{V_n - 2n\lambda(1-\lambda)}{2\sqrt{n}\lambda(1-\lambda)} < z\right) = \Phi(z). \tag{9}$$

This classical result leads to another classical randomness run tests.

It is worth noticing that the lengths of the longest run coincide under the classical and Feller's definitions. This is a useful characteristic for testing randomness. The limiting distribution of the longest run ν_n, exists only along the sequence with the fixed value of the fractional part of $\log_2 n$. More specifically, let $\{\log_2 n\}$ and $[\log_2 n]$ denote the fractional and integer parts of $\log_2 n$ respectively. Then

$$P\left(\nu_n - [\log_2 n] < k\right) \approx \exp\{-2^{-[k+1+\{\log_2 n\}]}\}. \tag{10}$$

For practical tests of randomness the distribution in (10) is not very convenient. A more practical test for a string of length n, such that that $n = MN$, is as presented in Section 2. If in the block (of size M) we have m ones (and $M-m$ zeroes), the conditional probability that the longest string of ones ν in this block is less than or equal to k has the following form with $U = \min\left(M-m+1, \left[\frac{m}{k+1}\right]\right)$ (see David and Barton (1962))

$$P\left(\nu \le k|m\right) = \frac{1}{\begin{pmatrix} M \\ m \end{pmatrix}} \sum_{j=0}^{U}(-1)^j \begin{pmatrix} M-m+1 \\ j \end{pmatrix}\begin{pmatrix} M-j(k+1) \\ M-m \end{pmatrix},$$

so that

$$P\left(\nu \le k\right) = \frac{1}{2^M}\sum_{r=0}^{M}\begin{pmatrix} M \\ r \end{pmatrix} P\left(\nu \le k|r\right), \tag{11}$$

The theoretical probabilities $\pi_0, \pi_1, \ldots, \pi_K$ of the classes are determined from (11).

5 Tests based on patterns

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) of a given length. The tests discussed here utilize the observed numbers of words which appear with a given frequency (i.e. which are missing, appear exactly once, exactly twice, etc.)

5.1 Tests based on the frequencies of patterns

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, $m = \log_2(n/\lambda)$. This assumption is needed for the following Poisson approximation to be applicable. An important role belongs to the set $\{j, 1 \le j \le m, i_{j+k} = i_k, k = 1, \ldots, m-j\}$, 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 ids 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 $W = W(m,n)$ this number of occurrences of the given nonperiodic pattern. The natural way to calculate W is as the sum,

$$W = \sum_{j=1}^{n-m+1} I(\epsilon_{j+k-1} = i_k, k = 1, \cdots, m). \tag{12}$$

Then $EW = (n-m+1)2^{-m}$. Under condition (6) W has a Poisson limiting distribution (7) (see Barbour, Holst and Janson, 1992, Sec 8.4). For patterns with a fixed length m, the limiting distribution of standardized statistic W in (12) is normal.

Note that the statistic (12) 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 $\imath$ in the string when $\imath$ has periods. As was discussed in Section 4, if $\tilde{W} = \tilde{W}(m,n)$ is the number of possibly overlapping appearances of a periodic word of length m in the string, then the asymptotic distribution of $\tilde{W}$ is the compound Poisson distribution (8). The corresponding probabilities can be expressed in terms of confluent hypergeometric function $\Phi =_1 F_1$ (see Johnson, Kotz, Kemp, 1992, pp 378-79).

If U denotes a random variable with distribution in (8), then for $u \ge 1$ with $\eta = \lambda/2$

$$P(U = u) = \frac{e^{-\eta}}{2^u} \sum_{\ell=1}^{u} \left(\begin{array}{c} u-1 \\ \ell-1 \end{array} \right) \frac{\eta^\ell}{\ell!} = \frac{\eta e^{-2\eta}}{2^u} \Phi(u+1, 2, \eta).$$

For example,

$$P(U = 0) = e^{-\eta},$$

$$P(U = 1) = \frac{\eta}{2} e^{-\eta},$$

$$P(U=2) = \frac{\eta e^{-\eta}}{8}\left[\eta + 2\right],$$

$$P(U=3) = \frac{\eta e^{-\eta}}{8}\left[\frac{\eta^2}{6} + \eta + 1\right],$$

$$P(U=4) = \frac{\eta e^{-\eta}}{16}\left[\frac{\eta^3}{24} + \frac{\eta^2}{2} + \frac{3\eta}{2} + 1\right].$$

As in Section 2, to use the corresponding tests the string is partitioned into N substrings and the empirical frequencies of occurrences of aperiodic or periodic patterns within each substring are conjoined by the χ^2-statistic. When $K = 5$ classes, i.e. $\{U = 0\}, \{U = 1\}, \cdots, \{U = 4\}$, $\{U \geq 5\}$, the theoretical probabilities $\pi_0, \pi_1, \ldots, \pi_5$ of these classes are found from the above formulas.

5.2 Tests of Randomness Based on the Number of Missing Words

The mathematical foundation for the test based on the number of missing m-words is contained in papers by Tikhomirova and Chistyakov (1997a, 1997b). As a matter of fact, the number of missing two-letter words is also employed in the so-called "OPSO Theory" introduced in Marsaglia and Zaman (1993) and used in the OPSO test of randomness in the Diehard Battery (Marsaglia (1996)). In this test one takes non-overlapping substrings formed by zeros and ones of given length p to represent the letters of the new alphabet, so that there are $q = 2^p$ new letters. In OPSO test one counts the number of two-letter patterns (the original substrings of length $2p$) which never occurred. (In the Diehard test $p = 10, q = 2^{10}$.) Our goal here is to give a different derivation of characteristics of this test.

5.2.1 Correlation polynomials and generating functions

We denote now by $\epsilon_1, \ldots, \epsilon_n$, a sequence of i.i.d. random variables each taking values in the finite set $\{1, \ldots, q\}$ such that $P(\epsilon_i = \ell) = p_\ell, \ell = 1, \ldots, q$. Thus, for any word $\imath = (i_1 \ldots i_m), P(\imath) = p_{i_1} \cdots p_{i_m}$. The situation when $p_\ell \equiv q^{-1}$ corresponds to the randomness hypothesis.

The following *correlation polynomial* of two patterns plays an important role in the study of the distribution of the numbers of missing words. Let $\imath = (i_1 \cdots i_m)$ and $\jmath = (j_1 \ldots j_m)$ be two patterns (words) of length m. Put

$$C_{\imath\jmath}(z) = \sum_{k=1}^{m} \delta_{(i_{m-k+1}\cdots i_m),(j_1\ldots j_k)} p_{j_{k+1}} \cdots p_{j_m} z^{k-1}. \tag{13}$$

If $\imath = \jmath$, then the correlation polynomial is referred to as the *correlation polynomial.* We denote it by $A_\imath(z) = C_{\imath\imath}(z)$ and by $\mathcal{A}(z)$ the *autocorrelation matrix,*

$$\mathcal{A}(z) = \begin{pmatrix} A_\imath(z) & C_{\imath\jmath}(z) \\ C_{\jmath\imath}(z) & A_\jmath(z) \end{pmatrix}.$$

A special role is played by *aperiodic* words $\imath$ of length m (discussed in Section 5.1), for which $A_\imath(z) = z^{m-1}$.

The first object of interest is the probability $\pi_\jmath(n)$ that a fixed pattern $\jmath$ is missing in the string of length n. According to Theorem 3.3 of Guibas and Odlyzko (1981) the probability generating function

$$F_\imath(z) = \sum_n \frac{\pi_\imath(n)}{z^n}$$

has the form

$$F_\imath(z) = \frac{zA_\imath(z)}{(z-1)A_\imath(z) + P(\imath)} = \frac{zA_\imath(z)}{P_q(z)}.$$

Then $P_q(z)$ is a polynomial of degree m,

$$P_q(z) = z^{m-1}(z-1) + \sum_{k=1}^{m-1} \delta_{(i_{m-k}\cdots i_m),(i_1\ldots i_k)} p_{i_{k+1}} \cdots p_{i_m} z^{k-1}(z-1) + P(\imath)$$

$$= \prod_{j=1}^{m} (z - z_j).$$

One obtains

$$F_\imath(z) = \sum_{k=1}^{m} \delta_{(i_{m-k}\cdots i_m),(i_1\ldots i_k)} p_{i_{k+1}} \cdots p_{i_m} \sum_{n=0}^{\infty} \frac{1}{z^{n+m-k}} \sum_{k_1+\cdots+k_m=n} z_1^{k_1} \cdots z_m^{k_m}.$$

Thus 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 $p_k \equiv 1/q$ and $\jmath$ is an aperiodic template,

$$\pi_\jmath(n) = \sum_{k_1+\cdots+k_m=n} z_1^{k_1} \cdots z_m^{k_m}. \tag{14}$$

The probabilities of the form, P (words $\imath$ and $\jmath$ are missing), also can be determined from the generating function the form of which follows from the simultaneous equations of Theorem 3.3, p 195 in Guibas and Odlyzko (1981). According to this theorem the generating function for probabilities of two specific words $\imath$ and $\jmath$ to be missing in the string of length n depends on the autocorrelation matrix $\mathcal{A}(z)$ in the following way

$$F_{\imath\jmath}(z) = \sum_n \frac{P(\text{words } \imath, \jmath \text{ are missing in } n\text{-string})}{z^n}$$

$$= \frac{z|\mathcal{A}(z)|}{(z-1)|\mathcal{A}(z)| + P(\jmath)A_\imath(z) + P(\imath)A_\jmath(z) - P(\imath)C_{\imath\jmath}(z) - P(\jmath)C_{\jmath\imath}(z)}. \tag{15}$$

Here $|\mathcal{A}(z)| = A_\imath(z)A_\jmath(z) - C_{\imath\jmath}(z)C_{\jmath\imath}(z)$ denotes the determinant of $\mathcal{A}(z)$. Note that in the simultaneous equations in Theorem 3.3 on p 195 of of Guibas and

Odlyzko (1981) the generating function $Q(z)$ must have factors $P(A), \dots, P(T)$ in all equations except the first one.

These formulas for the generating functions lead to the asymptotic behavior of the first two moments of the number of missing words. In the situation when $p_k \equiv 1/q$

$$A_\jmath(z) = z^{m-1} + \sum_{k=1}^{m-t} \delta_{(i_{m-k+1} \cdots i_m),(i_1 \dots i_m)} \frac{z^{k-1}}{q^{m-k}}.$$

Here $t, t \geq 1$, is the period of $\jmath$, i.e. the smallest positive integer for which $(i_{t+1} \cdots i_m) = (i_1 \dots i_{m-t})$. If $\jmath$ is an aperiodic template, $A_\jmath(z) = z^{m-1}$, then we put $t = \infty$.

It is not difficult to show that the largest root $\hat{z}$ of the equation $P_q(z) = 0$ is real, $\hat{z} < 1$, $\hat{z} \to 1$ as $q \to \infty$. Assume that $n/q^m \to a$ with a fixed positive a. Then the asymptotic approximation for the probabilities $\pi_\jmath(n)$ follows from the form of the generating function,

$$\pi_\jmath(n) = \pi_\jmath^t(n) \sim \frac{\hat{z}^{n+m-1}}{P_q'(\hat{z})} \left[1 + \frac{1}{\hat{z}^t q^t}\right].$$

One has

$$\hat{z} = 1 - \frac{1}{q^m} + \frac{1}{q^{m+t}} + o\left(\frac{1}{q^{m+r}}\right),$$

so that

$$\pi_\jmath^t(n) \sim e^{-a} \left[1 + \frac{a}{q^t}\right]. \tag{16}$$

For aperiodic words, the analysis of the polynomial equation

$$(z-1)z^{m-1} + q^{-m} = 0$$

shows that

$$\hat{z} = 1 - \frac{1}{q^m} - \frac{m-1}{q^{2m}} + O\left(\frac{1}{q^{2m}}\right).$$

It follows that in this case

$$\pi_\jmath^\infty(n) \sim e^{-a} \left[1 - \frac{(2m-1)a}{2q^m} + \frac{m-1}{q^m}\right]. \tag{17}$$

The form of the probabilities (17) and (16) leads to the formula for the expected value of the number of missing m-words, X. Let $q_t, t = 1, \dots, m-1, \infty$ denote the total number of words whose correlation polynomial has the form (13) (with $t = \infty$ corresponding to aperiodic words). Then

$$\sum q_t = q^m,$$

and as $q \to \infty$ for $t = 1, \dots, m-1$

$$\frac{q_t}{q^t} \to 1, \quad \frac{q_\infty}{q^m} \to 1.$$

With $\pi_\jmath^t(n)$ determined from (17) and (16), one gets

$$\mathbf{E}X = \sum_t q_t \pi_\jmath^t(n) = e^{-a} q^m + e^{-a}\left[m - 1 - \frac{a}{2}\right] + O\left(\frac{1}{q}\right). \tag{18}$$

This agrees from the formula from Tikhomirova and Chistyakov (1997a) on p 23 as these authors studied the number of missing words of length $m (= s)$ among first n words of length m, so that the total string length in their paper is $n + m - 1$. A similar identity (see the next Section) can be derived when $\pi_i = q^{-1} + q^{-3/2}\eta_i$ with $\sum \eta_i = 0$.

The derivation of the asymptotic formula for the variance is more cumbersome. To obtain this formula, one can use the fact that $X = \sum_\jmath x_\jmath$ where $x_\jmath$ is 0 or 1 according to occurrence of the word $\jmath$ in the string of length n. Thus, $Ex_\jmath x_\ell = P\,(\text{words } \jmath \text{ and } \ell \text{ are missing})$ and

$$\mathbf{Var}(X) = \sum_\jmath \mathbf{Var}(x_\jmath) + \sum_{\jmath \neq \ell} \mathbf{Cov}(x_\jmath, x_\ell)$$

$$= \sum_\jmath \pi_\jmath(n)[1 - \pi_\jmath(n)] + \sum_{\jmath \neq \ell} [P\,(\text{words } \jmath \text{ and } \ell \text{ are missing}) - \pi_\jmath(n)\pi_\ell(n)]. \tag{19}$$

The probabilities (16) and (17) have been determined, and the needed probabilities, $P\,(\text{words } \jmath \text{ and } \ell \text{ are missing})$, can be determined from the generating function (15). It turns out that the main contribution to this sum is due to pairs of aperiodic words such that at least one of the polynomials $C_{\imath\jmath}(z)$ or $C_{\jmath\imath}(z)$ vanishes.

Note that the analysis of the polynomial in the denominator of (15) corresponding to $C_{\imath\jmath}(z) = C_{\jmath\imath}(z) = 0$

$$(z-1)z^{m-1} + 2q^{-m} = 0$$

is the same as above. The probability of a pair of such words to be missing has the form

$$\pi_{\imath\jmath}(n) = e^{-2a}\left[1 - \frac{2(2m-1)a}{q^m} + \frac{2(m-1)}{q^m}\right] + O\left(\frac{1}{q^{2m}}\right).$$

If $C_{\imath\jmath}(z)$ has degree $m - 1 - u$ and $C_{\jmath\imath}(z) \equiv 0$, then the corresponding probability is of the form

$$\pi_{\imath\jmath}(n) = e^{-2a}\left[1 + \frac{a}{q^u}\right] + O\left(\frac{1}{q^{m \wedge 2u}}\right).$$

5.2.2 Two-letter words

When $m = 2$, the number of missing pairs has been used by Marsaglia and Zaman (1993) as a statistic for randomness testing. In this situation one obtains

$$\mathbf{Var}(X) = e^{-a}[1 - (1+a)e^{-a}]q^2 + a^2 e^{-2a} q$$

$$+e^{-a}\left[1-\frac{a}{2}-e^{-a}\left(1+2a-\frac{a^2}{2}-\frac{a^3}{2}\right)\right]+O\left(\frac{1}{q}\right). \tag{20}$$

The theoretical justification for approximate normality of the distribution of X for a fixed $\mathbf{s}$ when $n \to \infty, n/q^2 \sim a$ follows from a limit theorem by Mikhailov (1989) as the crucial condition there, $\mathbf{Var}(X) \to \infty$, is met.

To sum up, after the number X of missing two letter words in the string of length n has been evaluated, one finds the value $(X - \mathbf{E}X)/\sqrt{\mathbf{Var}(X)}$ with $\mathbf{E}X$ determined from (18) and $\mathbf{Var}(X)$ determined from (20) which leads to the P-value corresponding to the standard normal distribution.

The asymptotic power of this test against the discussed alternatives for which probability of the i th letter ($i = 1, \ldots, q$) is of the form $q^{-1} + \eta_i q^{-3/2}$ is determined by the ratio $a^2\left[e^a - 1 - a\right]^{-1/2}$ (see Tikhomirova and Chistyakov, 1997b). The largest possible value of this quantity (corresponding to the most powerful test against these alternatives) is attained when $a = a^\star = 3.594$.. This means that the best relationship between q and n is $n \approx 3.6q^2$.

5.3 Serial test and approximate entropy test

The (generalized) serial tests, as well as more general entropy based tests, represent a battery of procedures based on the testing the uniformity of the distributions of patterns of given lengths on the basis of their empirical entropies. Let $\omega_{i_1 \cdots i_m}$ denote the frequency of the pattern $(i_1, \cdots, i_m)$ in the "circularized" string of bits $(\epsilon_1, \ldots, \epsilon_n, \epsilon_1, \ldots, \epsilon_{m-1})$.

The covariance matrix of $\omega_{i_1 \cdots i_m}$ can be expressed in terms of the correlation polynomials (13) defined in Section 5.2.1. Let the matrix Σ_m be formed by $n^{-1}\mathbf{Cov}\left(\omega_{i_1 \cdots i_m}, \omega_{j_1 \cdots j_m}\right)$. Then its elements have the form

$$\sigma_{i_1 \cdots i_m j_1 \cdots j_m} = \frac{1}{q^m}\delta_{(i_1 \cdots i_m),(j_1 \cdots j_m)} - \frac{2m-1}{q^{2m}}$$

$$+\sum_{r=1}^{m-1}\left[\delta_{(i_{r+1} \cdots i_m),(j_1 \cdots j_{m-r})} + \delta_{(i_1 \cdots i_{m-r}),(j_{r+1} \cdots j_m)}\right]\frac{1}{q^{m+r}}$$

$$= \frac{1}{q^m}\delta_{(i_1 \cdots i_m),(j_1 \cdots j_m)} - \frac{2m-1}{q^{2m}} + \frac{C_{ij}(1) + C_{ji}(1)}{q^m}. \tag{21}$$

The rank of the matrix Σ_{m+1} is $q^{m+1} - q^m$, an one of its generalized inverses Σ^-_{m+1} has a remarkably simple form

$$\Sigma^-_{m+1} = q^{m+1}\mathbf{Q} = q^{m+1}\left[\mathbf{I}_{m+1} - q^{-1}\left(\mathbf{e}_1\mathbf{e}_1^T \oplus \cdots \oplus \mathbf{e}_1\mathbf{e}_1^T\right)\right]. \tag{22}$$

Here $\mathbf{I}_{m+1}$ denotes the identity matrix of size $q^{m+1} \times q^{m+1}$. Thus $\mathbf{Q}$ is merely the projection onto the orthogonal complement to the space spanned by the q^m vectors $\mathbf{e}_1 \oplus \mathbf{0} \oplus \cdots \oplus \mathbf{0}, \ldots, \mathbf{0} \oplus \cdots \oplus \mathbf{0} \oplus \mathbf{e}_1$.

The ϕ-entropy of a discrete random variable with the distribution given by probabilities $\pi_1, \dots, \pi_M$, is defined as

$$E(\pi_1, \dots, \pi_M) = \sum_{j=1}^{M} \pi_j \phi(\pi_j).$$

Here $\phi(u), 0 \leq u \leq 1$, is assumed to be continuously twice differentiable function. Commonly $\phi(1) = 0$ and ϕ is convex, in which case with $\phi(u) = \varphi(Mu), E$ becomes φ-information-type divergence between our distribution and the uniform one. When $M = q^m$ and the probability distribution is that of all m-templates, one may define the ϕ-uncertainty as $\sum \nu_{i_1 \cdots i_m} \phi(\nu_{i_1 \cdots i_m})$, where $\nu_{i_1 \cdots i_m} = \omega_{i_1 \cdots i_m}/n$ denotes the relative frequency of the template $(i_1, \cdots, i_m)$ in the augmented (circular) version of the original string.

As one needs a tractable limiting distribution for the tests statistics based on ϕ-entropy, we put

$$\Phi^{(m)} = a_m \sum_{i_1 \cdots i_m} \nu_{i_1 \cdots i_m} \phi(\nu_{i_1 \cdots i_m}), \tag{23}$$

where

$$a_m = \frac{q^m}{\phi'\left(\frac{1}{q^m}\right) + \frac{1}{2q^m}\phi''\left(\frac{1}{q^m}\right)}.$$

The general definition of *approximate ϕ-entropy AH* of order $m, m \geq 1$, is

$$AH(m) = \Phi^{(m)} - \Phi^{(m+1)}.$$

Large sample theory shows that the limiting distribution of

$$n\left[\Phi^{(m)} - \Phi^{(m+1)} - a_m \phi\left(\frac{1}{q^m}\right) + a_{m+1}\phi\left(\frac{1}{q^{m+1}}\right)\right]$$

coincides with that of $-q^{m+1}Z^T\mathbf{Q}Z$ (see Rukhin, 2000). Here $\mathbf{Q}$ is defined by (22) so that $q^{m+1}\mathbf{Q}$ is a generalized inverse of Σ_{m+1}. It is well known (see for example Theorem 9.2.2 in Rao and Mitra (1971)) that this distribution is the χ^2-distribution with the degrees of freedom equal to the rank of Σ_{m+1}. Therefore the centered sequence $nAH(m)$ asymptotically has the χ^2-distribution with $q^{m+1} - q^m$ degrees of freedom.

The classical Pearson's χ^2 statistic corresponds to $\phi(u) = u$. With $a_m = q^m$, $a_{m+1}\phi(q^{-m-1}) = a_m\phi(q^{-m})$ one has

$$\Phi^{(m)} = q^m \sum_{i_1 \cdots i_m} \nu_{i_1 \cdots i_m}^2 = 1 + \frac{\psi_m^2}{n}.$$

Here

$$\psi_m^2 = \sum_{i_1 \cdots i_m} \frac{(\omega_{i_1 \cdots i_m} - nq^{-m})^2}{nq^{-m}} = nq^m \sum_{i_1 \cdots i_m} \left(\nu_{i_1 \cdots i_m} - q^{-m}\right)^2$$

$$= nq^m \sum_{i_1 \cdots i_m} \nu^2_{i_1 \cdots i_m} - n,$$

is a χ^2-type statistic. (It is a common mistake to assume that ψ^2_m itself has the χ^2-distribution.)

The corresponding statistics are called generalized serial tests of randomness of a binary sequence (Good, 1953, Menezes, van Oorschot and Vanstone, 1997)

$$\nabla \psi^2_m = \psi^2_m - \psi^2_{m-1} = nAH(m-1).$$

Also

$$\nabla^2 \psi^2_m = \psi^2_m - 2\psi^2_{m-1} + \psi^2_{m-2}$$

can be used in randomness testing. (It is put here $\psi^2_0 = \psi^2_{-1} = 0$.) Indeed it follows that $\nabla \psi^2_m$ has approximately the χ^2-distribution with 2^{m-1} degrees of freedom and the limiting distribution of $\nabla^2 \psi^2_m$ can be shown to be the χ^2-distribution with 2^{m-2} degrees of freedom. The convergence of $\nabla \psi^2_m$ to χ^2-distribution was originally proven by Billingsley (1956).

Thus for small values of m, $m \leq 20$, one can find the corresponding P-values (and there are $2m$ of those) from the usual formulas. This result for $\nabla \psi^2_2$ and the usual counting of frequencies is given by Menezes, van Oorschot and Vanstone (1997) on p 181, formula (5.2), incorrectly, as $+1$ should be replaced by -1.

Another classical choice, $\phi(u) = -\log u$, with $a_m \equiv 2$, leads to the traditional Shannon entropy. In this case the statistic (23) forms the basis for the definition of approximate entropy investigated in a series of papers by S. Pincus and co-authors (Pincus and Singer, 1996, Pincus and Kalman, 1997).

A sequence was defined to be m-irregular (m-random) if its approximate entropy takes the largest possible value. Pincus and Kalman (1997) evaluated approximate entropies when $m = 0, 1, 2$ for binary and decimal expansions of $e, \pi, \sqrt{2}$ and $\sqrt{3}$ with the surprising conclusion that the expansion of $\sqrt{2}$ demonstrated more irregularity than that of e. However, they were unable to get the limiting distribution, which prevented this characteristic to be used as a randomness test.

6 Tests based on data compression

The tests discussed in this section are based on patterns suggested by the data themselves. The heuristic idea is that random sequences are those that cannot be compressed or those that are most complex. The discussed tests are based on statistics whose (approximate) distributions can be evaluated. The most interesting of those would be the missing test based on evaluation of Kolmogorov's complexity.

6.1 Lempel-Ziv Complexity Test

There are several variations on the Lempel-Ziv algorithm (1977). For a binary sequence the proposed test proceeds as follows:

1. Parse the sequence into consecutive disjoint words such that the next word is the shortest template not seen before.
2. Number the words consecutively (in base 2).
3. Assign each word a prefix and a suffix; the prefix is the number of the previous word that matches all but the last digit; the suffix is the last digit.

Note that it is possible, for small n, that the Lempel-Ziv compression is actually longer than the original representation. To see this more concretely, consider the sequence 010110010. It parses as $0, 1, 01, 10, 010$, giving five words. The first word has prefix 0 (since there is no previous word) and suffix 0; the second has prefix 0 and suffix 1, and so forth, giving $(0,0), (0,1), (1,1), (2,0), (3,0)$, which should be expressed in base 2 numbering as $(00,0), (00,1), (01,1), (10,0), (11,0)$. In principle, one can slightly extend this by adding a starting block of ones to indicate how many digits are needed for the prefix; this convention enables one to completely recover the entire sequence from the compressed sequence without any other knowledge. But such an extension is not pertinent to the testing issue.

Following the work of Aldous and Shields (1988), let $W(n)$ represent the number of words in the parsing of a binary random sequence of length n. They show that

$$\lim_{n\to\infty} \frac{E[W(n)]}{n/\log_2 n} = 1,$$

so that expected compression is asymptotically well-approximated by $n/\log_2 n$. Moreover,

$$P\left(\frac{W(n) - E[W(n)]}{\sigma[W(n)]} \leq w\right) \to \Phi(w).$$

However, Aldous and Shields were unable to determine the value of $\sigma[W(n)]$.

That difficulty was overcome by Kirschenhofer, Prodinger, and Szpankowski (1994) who prove that

$$\sigma^2[W(n)] \sim \frac{n[C + \delta(\log_2 n)]}{\log_2^3 n} \tag{24}$$

where $C = 0.26600$ (to five significant places) and $\delta(\cdot)$ is a slowly varying continuous function with mean zero and $|\delta(\cdot)| < 10^{-6}$.

To test a sequence for randomness compress this sequence using the Lempel-Ziv algorithm (as defined above). If the reduction is statistically significant when compared to the expected result, declare the sequence to be nonrandom. More exactly, given a sequence, parse it and count the number W of words obtained. It is not necessary to go through the complete Lempel-Ziv encoding since the test uses only the number of words. Then calculate the ratio

$$\frac{W - n/\log_2 n}{\sqrt{\frac{.266n}{\log_2^3 n}}}$$

and report the P-value corresponding to the two-sided alternative. (Some patterned sequences actually are flagged by being too long after compression.)

6.2 Maurer's "Universal Statistical" Test

The so-called "universal" test introduced by Maurer (1992) is the sole nonstandard test for randomness included in the *Handbook of Applied Cryptography* by Menezes et al (1997), and the sole compression-type randomness test in the package CRYPT-X, Gustafson et al (1994). Maurer calls this test universal in the sense that "it can detect any significant deviation of a [generator's] output statistics from the statistics of a truly random bit source when the [stream] can be modeled as [the output from] an ergodic stationary source with finite memory." Maurer's test statistic relates closely to the per-bit entropy of the stream, which allegedly is "the correct quality measure for a secret-key source in a cryptographic application."

The test is a compression-type test "based on the idea Ziv (1990) that a universal statistical test can be based on a universal source coding algorithm. A generator should pass the test if and only if its output sequence cannot be compressed significantly." According to Maurer, the source-coding algorithm due to Lempel-Ziv (1977) and discussed in Section 6.1 "seems to be less suited for application as a statistical test" because it seems to be difficult to determine the distribution of the test statistic.

The test requires a long (of the order $10 \cdot 2^L + 1000 \cdot 2^L$ with $6 \leq L \leq 16$) sequence of bits which it divides into two stretches of L-bit blocks: Q ($\geq 10 \cdot 2^L$) initialization blocks and K ($\approx 1000 \cdot 2^L$) test blocks. The order of magnitude of Q is specifically chosen to ensure that with a high probability all possible L-bit binary patterns do in fact occur in the initialization section. The test is ill suited for large values of L because the initialization takes time exponential in L. The parameter K represents the number of remaining blocks in the sequence being evaluated, it is not taken to be an input parameter since it is to be maximized. With $Q = 10(2^L)$ put $K = [n/L] - Q$.

The core of the test is to look back through the entire sequence while inspecting the test segment of L-bit blocks, checking for the nearest previous exact L-bit template 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), i.e.

$$f_n = \frac{1}{K} \sum_{i=Q+1}^{i=Q+K} \log_2(\#\text{indices since previous occurrence of ith template}).$$

The algorithm achieves this efficiently by subscripting of a dynamic look-up table making use of the integer representation of the binary bits constituting the template blocks. A P-value is obtained from the normal error function based on the standardized version of the statistic, with the test statistic's mean given by the formula (16) in Maurer (1992),

$$Ef_n = 2^{-L} \sum_{i=1}^{\infty} (1 - 2^{-L})^{i-1} \log_2 i.$$

In other terms the expected value of the test statistic f_n is that of the random variable $\log_2 G$ where $G = G_L$ is a geometric random variable with the parameter $1 - 2^{-L}$.

There are by now several versions of empirical approximate formulas for the variance of the form

$$Var(f_n) = c(L, K) Var(\log_2 G)/K.$$

Here $c(L, K)$ 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

$$c(L, K) = 0.7 - \frac{0.8}{L} + \left(1.6 + \frac{12.8}{L}\right) K^{-4/L}.$$

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." In other terms according to the Table 1.2 in Coron and Naccache (1998) the ratio of the standard deviation of f_n obtained from the approximation above to the true standard deviation considerably deviates from one.

In view of this fact and also since all approximations are based on the "admissible" assumption that $Q \to \infty$, one may 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. For this test the original sequence must be partitioned in a number r (say $r \le 20$) of substrings on each of which the value of the universal test statistic is evaluated (for the same value of parameters K, L and Q). The sample variance is evaluated then, and the P-value is obtained from the t-distribution with $r - 1$ degrees of freedom. Actually, the same observation is true with regard to the Lempel-Ziv compression test in Section 6.1, as the asymptotic formulas for the variance seem to underestimate finite-sample values of this characteristic.

Another intriguing possibility is to use the approximate two moments of the statistical estimate of entropy derived by Vatutin and Mikhailov (1995) in a similar problem.

6.3 Rank of random matrices test

The following is a test based on the result of Kovalenko (1972) according to which the linear rank R of $M \times Q$ random binary matrix takes values $r = 0, 1, 2, \ldots, m = \min(M, Q)$ with probabilities

$$p_r = 2^{r(Q+M-r)-MQ} \prod_{i=0}^{r-1} \frac{(1 - 2^{i-Q})(1 - 2^{i-M})}{1 - 2^{i-r}}.$$

It is suggested to implement this result for $M = Q \ge 10$. The number M is then a parameter of this test, so that ideally $n = M^2 N$ where N is the new "sample size". In practice we will look for values M and N such that the discarded part of the string, $n - NM^2$ is fairly small.

The rationale for this choice is that

$$p_M \approx \prod_{j=1}^{\infty} \left[1 - \frac{1}{2^j}\right] = 0.2888..,$$

$$p_{M-1} \approx 2p_M \approx 0.5776..,$$

$$p_{M-2} \approx \frac{4p_M}{9} \approx 0.1284..$$

and all other probabilities are very small (≤ 0.005) when $M \geq 10$.

For the obtained N square matrices evaluate their ranks $R_\ell, \ell = 1, \ldots, N$ and determine the frequencies F_M, F_{M-1} and $N - F_M - F_{M-1}$ of the values M, $M-1$ and of ranks not exceeding $M-2$,

$$F_M = \#\{R_\ell = M\},$$

$$F_{M-1} = \#\{R_\ell = M-1\}.$$

To apply χ^2-test use the classical statistic

$$\chi^2 = \frac{(F_M - 0.2888N)^2}{0.2888N} + \frac{(F_{M-1} - 0.5776N)^2}{0.5776N}$$

$$+\frac{(N - F_M - F_{M-1} - 0.1336N)^2}{0.13336N},$$

which under the null (randomness) hypothesis has the approximate χ^2-distribution with 2 degrees of freedom. The reported P-value is $\exp\{-\chi^2(obs)/2\}$.

6.4 Linear Complexity for Testing Randomness

This test pertains to an application of the notion of linear complexity which is related to one of the main components of many keystream generators, namely, Linear Feedback Shift Registers (LFSR). Such a register of length L 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. An LFSR generates a given binary sequence if this sequence is the output of the LFSR for some initial state. For a given sequence $S^n = (\epsilon_1, \ldots, \epsilon_n)$, its *linear complexity* $L(S^n)$ is defined as the length of the shortest LFSR that generates S^n as its 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 S^n is truly random, the formulas for the mean, $\mu_n = EL(S^n)$, and the variance, $\sigma_n^2 = Var(L(S^n))$, of the linear complexity $L(S^n) = L_n$ are well known. The computer package Crypt-X suggests that the distribution of the ratio $(L_n - \mu_n)/\sigma_n$ is close to that of a standard normal variable, so that the corresponding P-values of the test for randomness can be found from the normal error function. Indeed the paper by Gustafson et al (1994) p 693 claims that "for large n, $L(s^n)$ is approximately normally distributed with mean $n/2$ and a variance $86/81$ with the standard normal statistic $z = \left(L(s^n) - \frac{n}{2}\right)\sqrt{\frac{81}{86}}$." This fact is completely false. Even the mean value μ_n does not asymptotically behave precisely as $n/2$, and in view of boundedness of the variance, this difference becomes significant. More importantly, the tail probabilities of the limiting distribution are much larger than these of the standard normal distribution.

Strictly speaking 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 arising. 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 L_n, the linear complexity of a random binary string, $P\left(L_n = 0\right) = \frac{1}{2^n}$,

$$P\left(L_n = L\right) = \frac{2^{\min(2n-2L, 2L-1)}}{2^n} \quad L = 1, \dots, n. \tag{25}$$

Rueppel (1986), Proposition 4.2 derives the formula for the mean μ_n from these probabilities,

$$\mu_n = EL_n = \frac{1}{2^n}\Big[\sum_{j=1}^{m} j2^{2j-1} + \sum_{j=m+1}^{n} j2^{2n-2j}\Big]$$

$$= \frac{n}{2} + \frac{9+(-1)^{n+1}}{36} - \frac{1}{2^n}\left[\frac{n}{3} + \frac{2}{9}\right] = \frac{n}{2} + \frac{4+r_n}{18} - q_n.$$

Here $m = \left[\frac{n}{2}\right]$ is the integer part of $n/2$, $r_n = [1-(-1)^n]/2$ denotes the remainder of n divided by 2, and $q_n = 2^{-n}[n/3 + 2/9]$. Thus r_n is a bounded (non-converging, oscillating sequence) and $q_n \to 0$. Therefore we will ignore the term q_n in the following formula (27). One can also use (25) to obtain a formula for the variance σ_n^2, which turns out to be a bounded function of n.

By using (25), one can obtain the limiting distributions for $L_n - \mu_n$. Indeed the moment generating function has the form

$$M_n(t) = E\exp\{tL_n\} = \frac{1}{2^n}\left[1 + \sum_{j=1}^{m} e^{tj}2^{2j-1} + \sum_{j=m+1}^{n} e^{tj}2^{2n-2j}\right]$$

$$= \frac{1}{2^n}\left[\frac{1}{2} + \frac{e^{(m+1)(t+2\log 2)} - 1}{2(4e^t - 1)} + e^{tn}\frac{e^{(n-m)(-t+2\log 2)} - 1}{4e^{-t} - 1}\right]. \qquad (26)$$

As the variance is a bounded function of n, we look at the limiting distribution of $L_n - \mu_n$. To derive this distribution it suffices to find the limit of moment generating functions $M_n(t)e^{-t\mu_n} = E\exp\{t[L_n - \mu_n]\}$. One has

$$E\exp\{t[L_n - \mu_n]\} = \frac{e^{-t\mu_n}}{2^{n+1}} + \frac{e^{(m-\mu_n+1)(t+2\log 2)} - e^{-t\mu_n}}{2^{n+1}(4e^t - 1)}$$

$$+\frac{e^{(m-\mu_n)t+(n-m)2\log 2} - e^{(n-\mu_n)t}}{2^n(4e^{-t} - 1)}.$$

In our notation this identity takes the form

$$E\exp\{t[L_n - \mu_n]\} = \frac{e^{-\frac{n}{2}(t+2\log 2)}e^{-t(\frac{4+r_n}{18}-q_n)}\,[4e^t - 2]}{2(4e^t - 1)} - \frac{e^{\frac{n}{2}(t-2\log 2)}e^{-t(\frac{4+r_n}{18}-q_n)}}{4e^{-t} - 1}$$

$$+\frac{e^{t(\frac{7}{9}-\frac{5r_n}{9}+q_n)}2^{1-r_n}}{4e^t - 1} + \frac{e^{t(-\frac{2}{9}-\frac{5r_n}{9}+q_n)}2^{r_n}}{4e^{-t} - 1}.$$

The first two terms in this formula tend to 0 when $|t| < 2\log 2$, and to ∞ otherwise, whereas the last two terms remain bounded.

If n is an odd number, i.e. when $r_n = 1$, and $|t| < 2\log 2$, the limit of this sequence exists

$$\lim_{n\to\infty} M_n(t)e^{-t\mu_n} = M_V(t) = \frac{e^{\frac{2}{9}t}}{4e^t - 1} + \frac{2e^{-\frac{7}{9}t}}{4e^{-t} - 1},$$

and when n is even, i.e. $r_n = 0$, the limit takes the form

$$\lim_{n\to\infty} M_n(t)e^{-t\mu_n} = M_U(t) = \frac{2e^{\frac{7}{9}t}}{4e^t - 1} + \frac{e^{-\frac{2}{9}t}}{4e^{-t} - 1}.$$

As

$$M_V(t) = \frac{1}{4}e^{-\frac{7}{9}t}\left[1 - \frac{e^{-t}}{4}\right]^{-1} + \frac{1}{2}e^{\frac{2}{9}t}\left[1 - \frac{e^{t}}{4}\right]^{-1}$$

$$= \frac{1}{4}\sum_{\ell=0}^{\infty}\frac{e^{-\left(\ell+\frac{7}{9}\right)t}}{2^{2\ell}} + \frac{1}{2}\sum_{k=0}^{\infty}\frac{e^{\left(k+\frac{2}{9}\right)t}}{2^{2k}},$$

the corresponding random variable V takes possible values $v_\ell^- = -\ell - \frac{7}{9}, \ell = 0, 1, \ldots$ with probabilities $P(V = v_\ell^-) = \frac{1}{2^{2\ell+2}}$, and the values of the form $v_k^+ = k + \frac{2}{9}, k = 0, 1, \ldots$ with probabilities $P(V = v_k^+) = \frac{1}{2^{2k+1}}$. As it was to be expected, the function $M_V(t)$ is finite only for $|t| < 2\log 2$.

Similarly,

$$M_U(t) = \frac{1}{2}\sum_{\ell=0}^{\infty}\frac{e^{-\left(\ell+\frac{2}{9}\right)t}}{2^{2\ell}} + \frac{1}{4}\sum_{k=0}^{\infty}\frac{e^{\left(k+\frac{7}{9}\right)t}}{2^{2k}},$$

so that the possible values of U are $u_\ell^- = -\ell - \frac{2}{9}, \ell = 0, 1, \ldots$ with probabilities $P(U = u_\ell^-) = \frac{1}{2^{2\ell+1}}$, and the values of the form $u_k^+ = k + \frac{7}{9}, k = 0, 1, \ldots$ with probabilities $P(U = u_k^+) = \frac{1}{2^{2k+2}}$. Thus $-U$ has the same distribution as V.

Because of this fact and the inconvenience of having two different limiting distributions, we suggest to adapt the following sequence of statistics,

$$T_n = (-1)^n[L_n - \xi_n] + \frac{2}{9}. \tag{27}$$

Here

$$\xi_n = \frac{n}{2} + \frac{4 + r_n}{18}. \tag{28}$$

These statistics, which for any n can take only integer values from $-m$ to $m+1$, converge in distribution to the random variable T. This limiting distribution is skewed to the right. While $P(T = 0) = \frac{1}{2}$, for $k = 1, 2, \ldots$

$$P(T = k) = \frac{1}{2^{2k}}, \tag{29}$$

for $k = -1, -2, \ldots$

$$P(T = k) = \frac{1}{2^{2|k|+1}}. \tag{30}$$

One obtains from (29)

$$P(T \geq k > 0) = \frac{1}{3 \cdot 2^{2k-2}},$$

for $k < 0$ (30) shows that

$$P(T \leq k) = \frac{1}{3 \cdot 2^{2|k|-1}}.$$

Thus one can evaluate the P-value corresponding to the observed value T_{obs}, which for $\kappa = [|T_{\text{obs}}|] + 1$ has the form

$$\frac{1}{3 \cdot 2^{2\kappa-1}} + \frac{1}{3 \cdot 2^{2\kappa-2}} = \frac{1}{2^{2\kappa-1}}.$$

In view of the discrete nature of this distribution one uses the strategy described in Section 2 for a partitioned string. We note in conclusion that a similar test can be used when the string is formed by random variables with the uniform distribution over a finite set, say, $\{1, \ldots, q\}$. As in the case when $q = 2$, the following statistics,

$$T_n = (-1)^n[L_n - \xi_n] + \frac{q}{(q+1)^2} \tag{31}$$

with

$$\xi_n = \frac{n}{2} + \frac{2q + r_n(q-1)^2}{2(q+1)^2},$$

take only integer values from $-m$ to $m+1$, for any n. They converge in distribution to the random variable T such that $P(T=0)=\frac{q-1}{q}$, for a positive integer k

$$P(T=k)=\frac{q-1}{q^{2k}},$$

for $k<0$

$$P(T=k)=\frac{q-1}{q^{2|k|+1}}.$$

This distribution is also skewed to the right; its moment generating function has the form

$$M_T(t)=\frac{q-1}{q^2e^{-t}-1}+\frac{q(q-1)e^t}{q^2e^t-1}.$$

7 Independence of tests

The performance of the discussed tests can be checked by the Kolmogorov-Smirnov test of uniformity on the P-values obtained from the sequences. However, it requires us to assume that the sequences generated to test uniformity are sufficiently random. Intuitively some tests should give independent answers, e.g., the monobit test and a run test, that conditions on frequencies, should assess completely different aspects of randomness. For the other tests, such as the cusum test and the run test, the resulting P-values seem to be correlated.

To understand the dependencies between the tests in order to eliminate redundant tests, and to ensure that the tests in the suite are able to detect a reasonable range of patterned behaviors a factor analysis of the resulting P-values has been performed. More precisely, to assess independence, 300 sequences of binary pseudorandom digits were generated, each of length $n=1,000,000$. All ($k=61$) tests in the suite were used to test the randomness of those sequences. With p_{ij} denoting the P-value of test i on sequence j, the transformation $z_{ij}=\Phi^{-1}(p_{ij})$ leads to normally distributed variables, provided that p_{ij} are uniformly distributed. Let $\vec{z}_j$ be the vector of transformed P-values corresponding to the jth sequence. We performed a principal components analysis on the $\vec{z}_1,\ldots,\vec{z}_{300}$. Usually, a small number of components suffices to explain a great proportion of the variability, and the number of these components can be used to quantify the number of "dimensions" of non-randomness spanned by the suite tests. Our analysis extracted 61 factors, equal to the current number of tests in the suite. The first factor is the one that explains the largest variability. If many tests are correlated, their P-values will load highly on this factor, and the fraction of total variability explained by this factor will be large. The second factor explains the second largest proportion of variability, subject to the constraint that the second factor is orthogonal to the first, etc. The fractions corresponding to the first 50 factors for tests based on Blum-Blum-Shub sequences of length $1,000,000$ showed that there is no large redundancy among our tests.

Additionally, the correlation matrix formed from the $\vec{z}_1, \ldots, \vec{z}_{300}$, was constructed via a SAS program. The same conclusion was supported by the structure of these matrices. The degree of duplication among the tests seems to be very small.

It must be realized that for a large number of tests some of the P-values must be small even for perfectly random sequences.

8 Acknowledgement

This work has been supported by the NSA grant $\#MDA904 - 00 - 100033$. The author is grateful to Yu. V. Prokhorov and A.M. Zubkov for their interest.

9 References

1. Aldous, D. and Shields, P. A Diffusion Limit for a Class of Randomly-Growing Binary Trees. *Probability Theory and Related Fields*, **79**, 509-542, 1988.

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

3. Baron, M. and Rukhin, A. L. Distribution of the number of visits of a random walk. *Stochastic Models*, 15, 593–597, 1999.

4. Billingsley, P. Asymptotic distributions of two goodness of fit criteria. *Ann. Math. Statist.*, 27, 1123–1129, 1956.

5. Coron, J-S, and Naccache, D. An Accurate Evaluation of Maurer's Universal Test. Proceedings of SAC'98, Lecture Notes in Computer Science, Berlin: Springer-Verlag, 1998.

6. David, F. N., and Barton D. E. *Combinatorial Chance.* New York, Hafner Publishing Co, 1962,

7. Feller, W. *An Introduction to Probability Theory and Its Applications.* J. Wiley, 3rd Edition, New York, NY, 1968.

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

9. Good, I.J. The serial test for sampling numbers and other tests for randomness. *Proc. Cambridge Philos. Soc.*, 47, 276–284, 1953.

10. L. J. Guibas and A. M. Odlyzko. String overlaps, pattern matching and nontransitive games. *J. Combinatorial Theory*, 30, A, 183–208, 1981.

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

12. Johnson, N.L., Kotz, S., and Kemp, A. *Discrete Distributions.* J. Wiley, 2nd edition, New York, NY, 1996.

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

14. Kovalenko I. N. On the distribution of the linear rank of a random matrix, *Probab. Theory Appl.*, 17, 354–359, 1972.

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

16. Marsaglia, G. A Current View of Random Number Generation. In *Computer Science and Statistics: Proceedings of the Sixteenth Symposium on the Interface*, 3-10. Elsevier Science Pub., New York, 1985.

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

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

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

20. Menezes, A. J., van Oorschot, P. C., and Vanstone, S. A. *Handbook of Applied Cryptography*, CRC Press, Boca Raton, FL, 1997.

21. Mikhailov, V. G. Asymptotic normality of decomposable statistics from the frequencies of m-chains. *Discrete Math. Appl.*, 1, 335–347, 1991.

22. Pincus, S. and Kalman, R. E. Not all (possibly) "random" sequences are created equal. *Proc. Natl. Acad. Sci. USA*, 94, 3513–3518, 1997.

23. Pincus, S. and Singer, B. H. Randomness and degrees of irregularity. *Proc. Natl. Acad. Sci. USA*, 93, 2083–2088, 1996.

24. Rao, C. R. and Mitra, C. K. *Generalized Inverse of Matrices and its Applications.* J. Wiley, New York, 1971.

25. Revesz, P. *Random Walk in Random and Non-Random Environments.* World Scientific, Singapore, 1990.

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

27. Rukhin, A. L. Approximate entropy for testing randomness. *Journal of Applied Probability*, 37, 2000, 88-100.

28. 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.

29. Skorokhod, A. V. and Slobodenyuk, N. P. *Limit Theorems for Random Walks.* (in Russian), Kiev, Naukova Dumka, 1970.

30. M. I. Tikhomirova and V. P. Chistyakov. On asymptotics of the moments of missing s- patterns (in Russian). *Discrete Math. Appl*, 9:12–29, 1997a.

31. M. I. Tikhomirova and V. P. Chistyakov. On statistical tests based on missing s- patterns (in Russian). In *Trudy po Diskretnoi Matematike.* TVP Publishers, 265–278, 1997b.

32. Vatutin, V. A. and Mikhailov, V. G. Statistical estimation of entropy of discrete random variables with a large number of possible values. *Russian Math. Surveys*, 50, 121-132, 1995.

33. Ziv, J. Compression, tests for randomness and estimating the statistical model of an individual sequence. *Sequences* (ed. R. M. Capocelli), Berlin: Springer-Verlag, 1990.

34. Ziv, J. and Lempel, A. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23, 337-343, 1977.