

# TESTING RANDOMNESS

Andrew Rukhin

[rukhin@cam.nist.gov](mailto:rukhin@cam.nist.gov)

Statistical Engineering Division

National Institute of Standards and Technology

Building 820, Gaithersburg, MD 20899-0001

and Department of Mathematics & Statistics

University of Maryland, Baltimore County (UMBC)

Baltimore, MD 21250, USA

# Motivation

## Importance of Testing for Randomness

“Future of the science will belong to simulation-based modeling”

Thompson, 2000

Bad generators ruin studies

Wide use of public key cryptography

Need for good secure encryption algorithms

All such algorithms are based on a generator of (pseudo) random numbers;

Many (up to 75%) of common random generators fail (Karian and Dudewitz, 1999)

Testing of such generators for randomness is vital

A number of classic tests of randomness in Knuth (1989)

But some of these tests pass patently nonrandom sequences (Marsaglia, 1985)

# Statistical Test Suite

The most popular collection of tests for randomness, the Diehard Battery, demands fairly long strings (up to  $2^{24}$  bits).

A commercial product, CRYPT-X (Gustafson et al., 1994) includes a few tests for randomness.

A battery of statistical tests for randomness was developed by Computer Security Division and Statistical Engineering Division of the National Institute of Standards and Technology.

<http://crsc.nist.gov/rng/>

A. Rukhin, J. Soto, J. Nechvatal, M. Smid, E. Barker, S. Leigh, M. Levenson, M. Vangel, D. Banks, A. Heckert, J. Dray, S. Vo

“A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications”, NIST Special Publication 800-22, Department of Commerce, 2000.

# Applications

A. Rukhin “ Testing Randomness: a Suite of Statistical Procedures”,  
Theor. Probab. Appl, 2000 45, 111-132.

Immediate practical use:

this suite evaluated 15 algorithm proposals to replace the 56-bit DES

DATA ENCRYPTION STANDARD

Out of these 15 proposals the algorithm

RIJNDAEL was the winner, and is now AES

ADVANCED ENCRYPTION STANDARD

# The problem

The problem is to test that in a series of bits  $\epsilon_k, k = 1, 2, \dots, n$ , values 0 and 1 are each taken with probability  $1/2$  independently one from another.

1 0 1 0 1 0 0 0 1 1 1 1 0 1 0 1 0 1 0 0 1 1 1 0 1 0 0 0 0 0 0 1 1 0 1 1 1  
0 0 1 1 1 0 0 1 0 0 0 1 1 1 0 0 1 1 0 0 0 0 1 1 0 0 0 1 0 1 0 0 1 1 1 0 1  
0 1 1 0 1 0 0 1 1 0 1 1 0 0 0 0 0 1 1 1 1 1 1 0 1 1 0 0 0 1 1 0 0 1 0 0 0  
1 1 1 1

In some situations it is more convenient to have

$X_k = 2\epsilon_k - 1, k = 1, 2, \dots, n$ , with  $X_k$  are taking values  $+1$  or  $-1$ .

# P-values

There are about 16 different tests (statistics) in the Suite

About 50 for several parameter specifications

The result of each test is summarized by the P-value

P-value represents the probability of observing the value of the test statistic which is more extreme in the direction of non-randomness.

P-value measures the support for randomness hypothesis on the basis of a particular test

Under the randomness hypothesis P-value has a uniform distribution on  $[0, 1]$

# Random walk

$S_n = X_1 + \dots + X_n$  a random walk.

The most basic (monobit) test is that of the hypothesis that in a sequence of random variables  $X$ 's or  $\epsilon$ 's the probability of ones is  $1/2$ .

It is derived from the Central Limit Theorem. For positive  $z$

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

For the observed value  $|s(obs)| = |X_1 + \dots + X_n|/\sqrt{n}$ ,

$$P\left(\frac{|S_n|}{\sqrt{n}} \geq |s(obs)|\right) = 2[1 - \Phi(|s(obs)|)].$$

is the P-value, the probability of observing the value of  $S_n$  which is more extreme in the direction of the probability of ones different from  $1/2$ .

# Testing substrings

In many tests one can partition the original series into smaller substrings, apply the proposed tests on these substrings.

P-values can be combined:

$$-2 \sum_{i=1}^k \log(P_i) \sim \chi^2(2k)$$

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

A test based on the the number of visits of the random walk (Baron, Rukhin, 1999)

# Runs

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.

0, 0, 1, 0, 1, 1, 0, 0, 0, 0

has the following three runs of zeros: 0, 0; 0, and 0, 0, 0, 0.

Not the only possible definition of a run:

Von Mises (1964) does not allow a run to start or to end with no symbol at all

According to his definition there is only one run of zeros, namely, 0.

*Length* of a run, or a run of a given length

The classical definition declares the length of three runs of zeros above to be 2, 1 and 4 respectively.

# Feller Runs

Feller (1968): different definition of a run of length  $r$  (runs are non-overlapping, form recurrent events)

The advantage of Feller's definition: the moments of the occurrences of a run of length  $r$  admit a fairly simple generating function with explicit formulas for the mean

$$\mu = 2^{r+1} - 2$$

the variance

$$\sigma^2 = 2^{2(r+1)} - (2r + 1)2^{r+1} - 2$$

For a fixed  $r$  the number  $N_r$  of Feller defined runs of length  $r$  is approximately normal,

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

# Limit theorems for runs

A similar limit theorem holds for the classical definition of a run of a fixed length  $r$  (Mood, 1940).

If  $M_r$  denotes the total number of runs of length  $r$  then  $(M_r - EM_r) / \sqrt{Var(M_r)}$  approximately standard normal.

The situation changes when the length  $r$  is allowed to grow as  $n$  increases

$W = W(r, n)$  the number of Feller defined, non-overlapping runs of length  $r$ .

$$n, r \rightarrow \infty \quad \frac{n}{2^{r+1}} \rightarrow \lambda > 0,$$

$W$  has a Poisson limit

$$P(W = k) \rightarrow e^{-\lambda} \frac{\lambda^k}{k!} \quad k = 0, 1, \dots$$

# Poisson approximation

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

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

The conditional limiting distribution of the total number  $V_n$  of runs (in the classical setting) for the fixed proportion of ones  $\lambda$  is normal

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

(Gibbons, 1971).

The length of the longest run of ones or zeros is a relevant characteristic for testing randomness.

# Longest run

The limiting distribution of the longest run  $\nu_n$ , exists only along the sequence with the fixed value of the fractional part of  $\log_2 n$ .

$$P(\nu_n - [\log_2 n] < k) \approx \exp\{-2^{-[k+1+\{\log_2 n\}]}\},$$

$\{\log_2 n\}$  and  $[\log_2 n]$  denote the fractional and integer parts of  $\log_2 n$  respectively.

When  $n = 2047$  and the length of the longest run should be about 10,

$$P(\nu_{2047} \geq 14) \approx 1 - \exp(-1/16) = 0.06059.$$

This test is used in FIPS 140-1

When  $n = 20000$  (power-up tests)

$$P(\nu_{20000} \geq 34) \approx 1 - \exp(-7.816 * 10^{-7}) = 7.81610^{-7}.$$

# Implementation

For practical tests of randomness of a long string of length  $n$ ,  $n = MN$ , is to partition it into  $N$  substrings each of length  $M$ . For the test based on the length of the longest run of ones  $\nu_j$  within the  $j$ -th substring of size  $M$ ,  $K + 1$  classes (depending on  $M$ ) are chosen.

For each of these substrings the frequencies,  $\nu_0, \nu_1, \dots, \nu_K$ , of values of the longest run of ones within each of these substrings belonging to any of  $K + 1$  chosen classes,  $\nu_0 + \nu_1 + \dots + \nu_K = N$ , are evaluated.

The empirical frequencies  $\nu_i, i = 0, \dots, K$  are conjoined by the  $\chi^2$ -statistic

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

Under the randomness hypothesis approximate  $\chi^2$ -distribution with  $K$  degrees of freedom.

# $\chi^2$ -statistic

Large values of  $\nu_n$  tell us that the sequence has clusters of ones; the generation of “random” sequences by humans tends to lead to small values of  $\nu_n$ .

Von Mises (1964) reports studies of the German philosopher, K. Marbe, who investigated the birth records in four cities, each record containing 50,000 entries. The longest run was 17 entries of the same sex in a row.

The probability of observing  $\nu_{50,000} \geq 17$  is about 0.15, it is quite likely to observe such an event in a series of four independent trials. The probability that  $\nu_{200,000} \geq 17$  is about 0.97.

# Data compression

These tests 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 employed tests are based on statistics whose (approximate) distributions under randomness assumption can be evaluated.

Lempel-Ziv Complexity Test

Maurer's "Universal Statistical" Test

Binary Rank of Random Matrices (Diehard)

Linear Complexity for Testing Randomness

# Patterns

Most conventional pseudo random numbers generators, (the linear congruential generators, lagged-Fibonacci generators used in IMSL, C++, and other packages) tend to show patterning: deterministic recursive algorithms.

## Serial test and approximate entropy test

The (generalized) serial tests and entropy based tests, test uniformity of the distributions of patterns of given lengths on the basis of their empirical counterparts.

It is natural to employ statistical tests based on the occurrences of words (patterns or templates) of a given length, say,  $m$ .

# Setting

To utilize the observed frequencies of words which appear in a random text a prescribed number of times (i.e. which are missing, appear exactly once, exactly twice, etc.)

$q = 2^L$  to implement the test on the basis of a string of binary bits, take all substrings formed by zeros and ones of length  $L$  to represent the letters of the new alphabet and count the number of new  $m$ -letter patterns (the original non-overlapping consecutive substrings of length  $m * L$ ) with given frequencies.

$$P(\epsilon_i = k) = p_k, k = 1, \dots, q$$

$\epsilon_i$  independent

The probability of the word  $\iota = (i_1 \dots i_m)$ ,

$$P(\iota) = p_{i_1} \cdots p_{i_m}.$$

$p_k \equiv q^{-1}$  the randomness hypothesis ( $q = 2^M$ )

# Missing words

The probability for a given word  $\iota$  to appear in the string of length  $n$  exactly  $r$  times can be approximated by the Poisson probability of the value  $r$ , with the parameter  $nP(\iota)$ .

Less intuitive the covariance structure for several such random variables

Marsaglia used the number of missing ( $r = 0$ ) two-letter words ( $m = 2$ ) in the “OPSO Theory”

$q = 2^{10}$ ; the letters of the new alphabet all substrings formed by zeros and ones of length 10; count the number of new two-letter patterns (the original non-overlapping consecutive substrings of length 20), which never occurred.

A Poisson limit theorem for the number of occurrences of a given word (without small periods)

Barbour, Holst and Jensen (1992)

# Pattern Correlation Polynomials

With the alphabet consisting of  $q$  letters, let  $i = (i_1 \cdots i_m)$  and  $j = (j_1 \cdots j_m)$  be two patterns (words) of length  $m$ .

$$C_{ij}(z) = \sum_{k=1}^m \delta_{(i_{m-k+1} \cdots i_m), (j_1 \cdots j_k)} p_{j_{k+1}} \cdots p_{j_m} z^{k-1}$$

*correlation polynomial.*

Guibas and Odlyzko (1981)

A special role is played by *aperiodic* words  $i$  of length  $m$  for which  $C_{ii}(z) = z^{m-1}$ .

$C(z)$  (asymmetric) *correlation matrix*,

$$C(z) = \begin{pmatrix} C_{ii}(z) & C_{ij}(z) \\ C_{ji}(z) & C_{jj}(z) \end{pmatrix}.$$

Regnier and Szpankowski (1997), Szpankowski (2001). Andrew Rukhin, NIST, UMBC – p.20/33

# Pattern Correlation Polynomials

The probability  $\pi_i^r(n)$  that a fixed pattern  $i$  appears in the string of length  $n$  exactly  $r$  times,

Guibas and Odlyzko (1981) when  $r = 0$ .

The probability generating function

$$F_i^0(z) = \sum_n \frac{\pi_i^0(n)}{z^n} = \frac{zC_i(z)}{B(z)},$$

$$B(z) = (z - 1)C_i(z) + P(i)$$

a polynomial of degree  $m$ .

# Waiting times

The expected waiting time until pattern  $\iota$  appears

$$\sum_n P(\iota \text{ did not appear in the first } n \text{ letters}) = F_\iota^0(1) = \frac{C_\iota(1)}{P(\iota)}$$

The more periods  $\iota$  has, the longer is the waiting time till it appears.

The odds of  $j$  to precede  $\iota$

$$\frac{C_\iota(1) - C_{\iota j}(1)}{C_{j j}(1) - C_{j \iota}(1)}.$$

## Penney Game

First player  $\iota = (100)$ ; Second player  $j = (110)$ .

$$C_\iota(z) = C_{j j}(z) = z^2, C_{j \iota}(z) = p_0 z, C_{\iota j}(z) = 0$$

The odds for the first player to win are only  $(1 - p_0) : 1$ .

# Generating functions

Typically,  $B(z)$  has  $m$  distinct roots, a unique largest positive root  $z_1 = \bar{x} < 1$ , all other roots  $z_k, k = 2, \dots$  lying in the circle of smaller radius  $\rho < \bar{x}$ .

$$\begin{aligned}\pi_i^0(n) &= \sum_{\ell=1}^m \frac{z_\ell^{n+m-1}}{\prod_{k:k \neq \ell}^m (z_\ell - z_k)} = \sum_{\ell=1}^m \frac{z_\ell^{n+m-1}}{B'(z_\ell)} \\ &= \frac{\bar{x}^{n+m-1}}{B'(\bar{x})} + O((\rho + \Delta)^n).\end{aligned}$$

For  $r \geq 1$ , a formula for the generating function  $F_i^r(z)$  in Fudos, Pitoura and Szpankowski (1995),

$$F_i^r(z) = \frac{z^m P(i) [(z-1)(C_n(z) - z^{m-1}) + P(i)]^{r-1}}{[B(z)]^{r+1}}.$$

# Setting

The generating function for the probabilities,  $\pi_{ij}^{rt}(n)$ , that a given word  $i$  occurs in the string of length  $n$  exactly  $r$  times and a word  $j$  occurs  $t$  times in Rukhin (2002).

Extensions to Markov sequences Rukhin (2006).

The asymptotic distribution of the number of words appearing a given number of times

$$n \rightarrow \infty, q \rightarrow \infty \quad n/q^m \rightarrow \alpha > 0$$

Assume the distribution of the alphabet letters is close to the uniform,

$$p_k = q^{-1} + q^{-3/2} \eta_k, \quad k = 1, \dots, q \quad \sum_{k=1}^q \eta_k = 0$$

$$\frac{1}{q} \sum_k \eta_k^2 \rightarrow \mathbf{B} > 0$$

# Expected values

$X^r = X_n^r$  the number of  $m$ -words, which occur exactly  $r$  times in a string of length  $n$

For  $r = 0, 1, \dots$

$$\mathbf{E}X^r = \sum_i \pi_i^r(n) = \frac{\alpha^r e^{-\alpha}}{r!} q^m \left[ 1 + \frac{m\mathbf{B}}{2} [\alpha^2 - 2\alpha r + r(r-1)] + O\left(\frac{1}{q^{3/2}}\right) \right].$$

When  $\eta_k \equiv 0$ ,

$$\mathbf{E}X^r = \frac{\alpha^r e^{-\alpha}}{r!} \left[ q^m - \frac{\alpha}{2} + m + r - 1 - \frac{r(2mr + 4m - r - 5)}{2\alpha} \right] + O\left(\frac{1}{q}\right).$$

# Variance

For  $r \neq t$

$$\text{Cov}(X^r, X^t) = -q^m \frac{e^{-2\alpha} \alpha^{r+t}}{r!t!} \left[ \alpha - r - t + 1 + \frac{rt}{\alpha} \right] + O\left(q^{m-1}\right),$$

a similar formula for  $\text{Var}(X^r)$

Kolchin, Sevastyanov and Chistyakov (1978): formulas for the first two moments of the joint distribution of the number of words appearing a prescribed number of times when the occurrences of these words are independent, i.e. when the words appearances in the non-overlapping  $m$ -blocks are counted.

A surprising fact: the asymptotic behavior of the expected value and of the covariance matrix is the same for overlapping and non-overlapping occurrences.

# Optimal test

The goal: to combine statistics  $X^0, X^1, \dots, X^R$  into one test.

When  $n \rightarrow \infty$ ,  $n/q^m \rightarrow \alpha$  with a fixed positive  $\alpha$ , the asymptotic joint distribution of the random variables  $X^0, X^1, \dots, X^R$  is normal with the covariance matrix  $\Xi$ .

$p_r(\alpha) = \frac{\alpha^r e^{-\alpha}}{r!}$  Poisson probabilities

The elements of matrix  $\Xi$

$$\xi_{rr} = p_r(\alpha) \left[ 1 - p_r(\alpha) \left( \alpha - 2r + 1 + \frac{r^2}{\alpha} \right) \right],$$

$$\xi_{rt} = -p_r(\alpha)p_t(\alpha) \left[ 1 + \frac{(\alpha - r)(\alpha - t)}{\alpha} \right],$$

for  $r \neq t$

# Optimal test

The optimal test of the null hypothesis  $H_0 : \eta_i \equiv 0$  within the class of linear statistics

$$S = \sum_{r=0}^R w_r X^r$$

$S$  is asymptotically normal both under the null hypothesis and the alternative  $H_1 : B > 0$ .

The Pitman efficiency of this statistic is determined by its efficacy, the distance between the means under the null hypothesis and the alternative, divided by the standard deviation (common to the null hypothesis and the alternative),

$$\text{eff}(S) = \frac{m \left| \sum_{r=0}^R w_r p_r(\alpha) [\alpha^2 - 2\alpha r + r(r-1)] \right|}{2 \left( \sum_{r,t} \xi_{rt} w_r w_t \right)^{1/2}}$$

# Optimal test

$(R + 1)$ -dimensional vector  $\mathbf{w}$  has coordinates  $w_0, \dots, w_R$

$\mathbf{b}$  has coordinates

$$p_r(\alpha)(\alpha^2 - 2\alpha r + r(r - 1))$$

Maximization of this ratio leads to the solution

$$\mathbf{w} = \Xi^{-1}\mathbf{b}$$

With this optimal choice of the weights,

$$\text{eff}(S) = m \frac{\sqrt{\mathbf{b}^T \Xi^{-1} \mathbf{b}}}{2}.$$

The test based on  $S$  can be shown to be asymptotically optimal (not only within the class of linear statistics.)

# Two-letter words

The approximations to the expected value and the variance of the number of missing pairs.

These approximations form a basis for the distribution of a randomness testing statistic.

Tikhomirova and Chistyakov (1997)

Marsaglia and Zaman (1993)

The expected value of the number of missing words,  $X^0$ ,

$$EX^0 \approx e^{-\alpha} q^2 + e^{-\alpha} \alpha^2 \mathbf{B}q$$

$$+ e^{-\alpha} \left[ 1 - \frac{\alpha}{2} \right] + 4e^{-\alpha} \alpha^2 \mathbf{B}^2 [2 - 4\alpha + \alpha^2]$$

$$\text{Var}(X^0) \approx e^{-\alpha} [1 - (1 + \alpha)e^{-\alpha}] q^2 + \alpha^2 e^{-2\alpha} q$$

$$+ e^{-\alpha} \left[ 1 - \alpha - e^{-\alpha} \left( 2 - 8\alpha - \frac{3\alpha^2}{2} \right) \right] \mathbf{B}q$$

# Two-letter words

The efficacy of the test based on the number of missing words is determined by

$$\alpha^2 [e^\alpha - 1 - \alpha]^{-1/2}$$

The largest possible value of this quantity (corresponding to the most powerful test when  $R = 0$ )

$$\alpha = \alpha^* = 3.594..$$

The best relationship between  $q$  and  $n$ , when  $m = 2$ , is  $n \approx 3.6q^2$ .

# $\chi^2$ -statistic

$\omega_i$  frequency of  $i$

The quadratic statistic

$$\chi^2 = \sum_i \frac{(\omega_i - nP(i))^2}{nP(i)} - \sum_{i_1 \dots i_{m-1}} \frac{(\omega_{i_1 \dots i_{m-1}} - p_{i_1} \dots p_{i_{m-1}})^2}{np_{i_1} \dots p_{i_{m-1}}}$$

asymptotically has  $\chi^2$ -distribution with  $q^m - q^{m-1}$  degrees of freedom.

A test of the null hypothesis according to which probabilities of letters in a random text are given numbers  $p_1, \dots, p_q$ . Reject the null hypothesis for large values of  $\chi^2$ -statistic; generalizes the well known serial test used for testing uniformity (Good, 1954).

# Conclusions

In view of importance of randomness testing, new stringent tests are desired.

The talk reviewed some of such tests.

Future work: tests based on two-dimensional patterns

An open problem: to develop a practical randomness test based on evaluation of Kolmogorov's complexity