Review of results on joint distributions of NIST STS tests and recommendations for choosing parameters of these tests

Savelov M.P.

Lomonosov Moscow State University

Relevance of the topic

Random number generators (RNG) are used in

- numerical methods (Monte Carlo method)
- modeling (models of physical experiments, computer games)
- programming (examples in «The Art of Programming» by D. Knuth)
- in cryptography (key generation)

Robert Coveyou: "The generation of random numbers is too important to be left to chance"

RNG quality

How to check the quality of the RNG?
The most famous early works: M. G. Kendall, B. Babington-Smith (1938, 1939).
An example of testing a specific RNG: I. M. Sobol, 1958.

Packages of statistical tests: NIST, Diehard, TestU01, Crypt-X, SPRNG, etc. Review: P. L'Ecuyer, 2017.

The most popular package of tests for testing random binary sequence generators: Rukhin A., Soto J., Nechvatal J., Smid M., Barker E., Leigh S., Levenson M., Vangel M., Banks D., Heckert A. , Dray J., Vo S. «A statistical test suite for random and pseudorandom number generators for cryptographic applications», NIST Special Publication 800-22 Revision 1a, ed. L. E. Bassham III, NIST, April 2010

About NIST STS

A. L. Rukhin, "Testing randomness: a suite of statistical procedures", *Probability theory. and its applications.*, **45**:1 (2000), 137–162.

A sequence $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ of zeros and ones is given.
Hypothesis H_0: «ε_i, $1 \le i \le n$, — i.i.d., with distribution $Bern(\frac{1}{2})$».
NIST: 15 statistical tests. For example:

- 1) «Monobit test»
$$T_{mon} = \frac{2S_n - n}{\sqrt{n}}.$$
- 2) «Frequency Test within a Block». Let N be fixed, put $M = \left[\frac{n}{N}\right]$, $\pi_j = \frac{1}{M}\sum_{i=(j-1)M+1}^{jM} \varepsilon_i \quad (1 \le j \le N)$ and
$$T_{fr} = 4M \sum_{j=1}^{N} \left(\pi_j - \frac{1}{2}\right)^2$$

About NIST STS

Other tests: «Discrete Fourier Transform (Spectral) Test», «Approximate Entropy Test» and so on.
Problem: statistics of 15 tests of NIST STS are not independent: A.M. Zubkov, A.A. Serov (2021); P. Burciu, E. Simion (2019) and others.

Our goals:

- Find the limiting joint distributions of various sets of statistics of the NIST STS tests;
- Select parameters in these tests such that corresponding statistics become asymptotically independent.

NIST: «...the tests should be independent of each other as much as possible».

Discrete Fourier Transform (Spectral) Test

«Discrete Fourier Transform (Spectral) Test». Put $X_k = 2\varepsilon_k - 1$ and $f_j = \sum_{k=1}^{n} X_k e^{\frac{2\pi i(k-1)j}{n}}, \quad j = 0, 1, \ldots, n-1.$
Fix $\gamma > 0$ and $c > 0$. Consider

$$T_{Fourier}(c, \gamma) = \frac{1}{\sqrt{\frac{n}{\gamma} \cdot 0,95 \cdot 0,05}} \left(\sum_{j=0}^{[\frac{n}{2}]-1} I_{|f_j| < \sqrt{cn}} - 0,95 \cdot \frac{n}{2} \right).$$

In old version of NIST STS $T_{Fourier}(3, 2)$ was used and it was supposed that $T_{Fourier}(3, 2) \xrightarrow{d} \mathcal{N}(0, 1)$.
Idea: $n^{-\frac{1}{2}}|f_j|$ are «almost independent» and $\mathcal{L}(n^{-\frac{1}{2}}|f_j|) \approx \sqrt{\mathrm{Exp}(1)} \Rightarrow$ indicators $I_{n^{-\frac{1}{2}}|f_j| < \sqrt{cn}}$ are «almost i.i.d.»,

$$c \approx (\text{the quantile of } Exp(1) \text{ of order } 0.95) = \ln 20 = 2.9957\ldots \approx 3,$$

$$\text{number of terms} = \frac{n}{2} \Rightarrow \gamma = 2.$$

Discrete Fourier Transform (Spectral) Test

$$T_{Fourier}(c,\gamma) = \frac{1}{\sqrt{\frac{n}{\gamma}\cdot 0,95\cdot 0,05}}\left(\sum_{j=0}^{[\frac{n}{2}]-1} I_{|f_j|<\sqrt{cn}} - 0,95\cdot\frac{n}{2}\right).$$

In old version of NIST STS $T_{Fourier}(3,2)$ was used and it was supposed that $T_{Fourier}(3,2) \xrightarrow{d} \mathcal{N}(0,1)$.
In new version of NISTS STS: $T_{Fourier}(\ln 20,4)$ is used and it's supposed that $T_{Fourier}(\ln 20,4) \xrightarrow{d} \mathcal{N}(0,1)$ (S. Kim, K, Umeno, A. Hasegawa, 2004).
Pareschi F., Rovatti R., Setti G. *IEEE Trans. Inf. Forensics Secur.* (2012): noticed that $T_{Fourier}(\ln 20,3.8) \xrightarrow{d} \mathcal{N}(0,1)$
The exact limiting distribution of $T_{Fourier}(\ln 20,\gamma)$ is unknown.
Hypothesis (Savelov, 2021): $T_{Fourier}(\ln 20,\gamma) \xrightarrow{d} \mathcal{N}(0,1) \Longleftrightarrow \gamma = \gamma^*$, where

$$\gamma^* = 2\left(1 - \frac{0,05\ln^2 0,05}{0,95}\right)^{-1} = 3.79030132194714...$$

Fix $c > 0$. For all $\varepsilon \in (0, c)$ fix a function $\phi_{c,\varepsilon} : \mathbb{R} \to \mathbb{R}$, such that

- $\phi_{c,\varepsilon}(x) = 1$ for $x < c - \varepsilon$ and $\phi_{c,\varepsilon}(x) = 0$ for $x > c + \varepsilon$,
- $\phi_{c,\varepsilon}(x) \in [0, 1]$ for $x \in \mathbb{R}$,
- $\phi_{c,\varepsilon}(x) \in C^2(\mathbb{R})$.

Recall:

$$T_{Fourier}(c, \gamma) = \frac{1}{\sqrt{\frac{n}{\gamma} \cdot 0,95 \cdot 0,05}} \left(\sum_{j=0}^{[\frac{n}{2}]-1} I_{n^{-1}|f_j|^2 < c} - 0,95 \cdot \frac{n}{2} \right).$$

G. Fay, P. Soulier (2001) $\Rightarrow$ we may prove the next theorem.

Theorem (2021)

For fixed $\varepsilon \in (0, \frac{1}{10}]$ *the statistics*

$$T^{\varepsilon}_{Fourier} = \frac{1}{\sqrt{\frac{n}{\gamma(\varepsilon)} \cdot 0,95 \cdot 0,05}} \left(\sum_{j=0}^{[\frac{n}{2}]-1} \phi_{\ln 20,\varepsilon}(n^{-1}|f_j^2|) - a(\phi_{\ln 20,\varepsilon}) \cdot \frac{n}{2} \right)$$

is well-defined and it's distribution converges to $N(0,1)$ *as* $n \to \infty$, *and, moreover,* $a(\phi_{\ln 20,\varepsilon}) \to 0.95$ *and* $\gamma(\varepsilon) \to \gamma^*$ *as* $\varepsilon \to +0$.

$$T_{Fourier}(c,\gamma) = \frac{1}{\sqrt{\frac{n}{\gamma} \cdot 0,95 \cdot 0,05}} \left(\sum_{j=0}^{[\frac{n}{2}]-1} I_{n^{-1}|f_j|^2<c} - 0,95 \cdot \frac{n}{2} \right).$$

Hypothesis: $T_{Fourier}(\ln 20, \gamma) \xrightarrow{d} \mathcal{N}(0,1) \Longleftrightarrow \gamma = \gamma^*$, where

$$\gamma^* = 2\left(1 - \frac{0,05 \ln^2 0,05}{0,95}\right)^{-1} = 3.79030132194714...$$

Recommendation:

It's better to use $T^{\varepsilon}_{Fourier}$ or $T_{Fourier}(\ln 20, \gamma^*)$ instead of $T_{Fourier}(\ln 20, 4)$ in Discrete Fourier Transform (Spectral) Test.

15 tests of NIST STS

1	Frequency (Monobit) Test
2	Frequency Test within a Block
3	Runs Test
4	Test for the Longest Run of Ones in a Block
5	Binary Matrix Rank Test
6	Discrete Fourier Transform (Specral) Test
7	Non-Overlapping Template Matching Test
8	Overlapping Template Matching Test
9	Maurer's "Universal Statistical" Test
10	Linear Complexity Test
11	Serial Test
12	Approximate Entropy Test
13	Cumulative Sums (Cusum) Test
14	Random Excursions Test
15	Random Excursions Variant Test

Idea 1: correlations of indicators

Let the i-th test, based on T_i, has the form $\{T_i > u_i\}$, where u_i is the quantile of T_i of order 0.99. We use a sample estimate of covariance of the values $I_{T_i>u_i}$ and $I_{T_j>u_j}$. Note:

$$\mathrm{cov}(I_{T_i>u_i}, I_{T_j>u_j}) = \mathbf{P}(T_i > u_i, T_j > u_j) - \mathbf{P}(T_i > u_i)\mathbf{P}(T_j > u_j).$$

Simion E., Burciu P. «A note on the correlations between nist cryptographic statistical tests suite» (2019): tests numbered 1, 11a, 11b, 13a, 13b, 14, 15

$$(1, 13a), (1, 13b), (13a, 13b), (14, 15), (11a, 11b).$$

Idea 2: p-value correlations

Let p_i be the p-value corresponding to the i-th test.
Correlation of p_i and p_j:

$$r_{ij} = \frac{\mathrm{cov}(p_i, p_j)}{\sqrt{\mathbf{D}p_i \mathbf{D}p_j}}.$$

$\hat{r}_{ij}$ is the sample Pearson correlation coefficient. We compare $\hat{r}_{ij}$ with zero.
Burciu P., Simion E., "A Systematic approach of NIST statistical tests dependencies" (2019): tests 1, 13a, 13b, 14, 15, 11a, 11b and

$$\hat{r}_{1,13a} = 0.745, \quad \hat{r}_{1,13b} = 0.753, \quad \hat{r}_{13a,13b} = 0.711, \quad \hat{r}_{11a,11b} = 0.679.$$

Pairs of dependent tests:

$$(1, 13a), (1, 13b), (13a, 13b), (11a, 11b).$$

Doganaksoy A., Sulak F., Uguz M., Şeker O., Akcengiz Z., Mutual correlation of NIST statistical randomness tests and comparison of their sensitivities on transformed sequences (2017)

Idea 3: distribution of $p_i - p_j$
If the i-th and j-th tests are independent, then $p_i - p_j \in [-1, 1]$ and

$$p_i - p_j \sim U_1 - U_2,$$

where U_1, U_2 are independent r.v. with uniform distribution on $[0, 1]$.
If the empirical distribution of $p_i - p_j$ differs from the distribution of $U_1 - U_2$, then statistics are dependent.
Fan L., Chen H., Gao S. A general method to evaluate the correlation of randomness tests (2014).
Pairs of dependent tests:

$$(1, 12), (1, 13), (3, 11), (3, 12), (11, 12), (12, 13).$$

Idea 4: Fail-Fail Ratio

If the i-th and j-th tests are independent, then

$$\mathbf{P}(T_j > u_j | T_i > u_i) = \mathbf{P}(T_j > u_j) = 0.01.$$

Condider L independent Bernoulli sequences ($Bern(\frac{1}{2})$) of length n. Approximately $\frac{L}{100}$ of them will fail the i-th test, and approximately $\frac{L}{10000}$ sequences will fail both tests.
We divide the number of sequences that failed both tests by the number of sequences that failed the i-th test and get «fail-fail ratio» (FFR).
$FFR \approx \mathbf{P}(T_j > u_j | T_i > u_i) = 0.01.$

Doğanaksoy A., Ege B., Muş K. (2008),
Sulak F., Uğuz M., Koçak O., Doğanaksoy A. (2017)

Idea 5: mutual information

Let X and Y be random variables. Let $H(X)$ be the entropy of the random vector X. $H(Y)$ and $H(X, Y)$ are understood similarly. Mutual information

$$I(X, Y) = H(X) + H(Y) - H(X, Y)$$

It is known: $I(X, Y) \geq 0$, and $I(X, Y) = 0 \Leftrightarrow X$ and Y are independent.
If $X = T_i$, $Y = T_j$ and T_i and T_j are independent, then $I(T_i, T_j) = 0$.
If $X = p_i$, $Y = p_j$ and T_i and T_j are independent, then $I(p_i, p_j) = 0$.
Karell-Albo J.A., Legòn-Pèrez C.M., Madarro-Capò E.J., Rojas O., Sosa-Gòmez G., (2020).

$$(1,3),\ (1,13a),\ (1,13b),\ (1,14),\ (1,15),\ (3,13a),\ (3,13b),\ (4,8),$$
$$(7,13b),\ (11a,11b),\ (11a,12),\ (13a,13b),\ (13a,14),$$
$$(13a,15),\ (13b,14),\ (13b,15),\ (14,15).$$

The main idea: $\mathrm{cov}(X, Y) = 0 \not\Rightarrow X$ and Y are independent, but $I(X, Y) = 0 \Rightarrow X$ and Y are independent.

Idea 6: Binomial distribution

Consider r tests. Suppose that each of them has error probability α. Let τ be the number of tests that reject H_0. Obviously, $0 \leq \tau \leq r$.

If all r statistics are independent then $\tau \sim Bin(\alpha, r)$ for H_0.

The difference between the distribution of τ and $Bin(\alpha, r)$ indicates that not all r tests are independent.

A. M. Zubkov, A. A. Serov, Mat. Vopr. Kriptogr., (2021).

Let $\alpha = 0.01$ and $r = 14$. Consider all tests from 1-st to 13-th inclusive, not counting the 7-th (14 in total). The random variable τ has a distribution different from $Bin(\alpha, r)$.

If we exclude from the previous set (with 14 tests) one of the statistics of the Serial Test and both statistics of the Cumulative Sums Test, then τ has a distribution that is not very different from $Bin(\alpha, r)$.

Theorems about statistics of NIST STS tests

- 1) «Monobit test»

$$T_{mon} = \frac{2S_n - n}{\sqrt{n}}.$$

- 2) «Frequency Test within a Block». Let N be fixed, put $M = \left[\frac{n}{N}\right]$, $\pi_j = \frac{1}{M}\sum_{i=(j-1)M+1}^{jM} \varepsilon_i \quad (1 \le j \le N)$ and

$$T_{fr} = 4M \sum_{j=1}^{N} \left(\pi_j - \frac{1}{2}\right)^2$$

Statistics of NIST STS tests

- 3) «Runs test»
  Put $V_n = 1 + \sum_{i=1}^{n-1} I_{\varepsilon_i \neq \varepsilon_{i+1}}$, $\pi = \frac{S_n}{n}$ и

$$T_{runs} = \frac{V_n - 2n\pi(1-\pi)}{2\sqrt{n}\pi(1-\pi)}.$$

- 4) «Non-overlapping Template Matching Test».
  As m-bit template B is a sting of ones and zeros of length m:
  $B = (b_1, b_2, \ldots, b_m) \in \{0,1\}^m$,

$$W_j = \sum_{i=(j-1)M+1}^{jM-m+1} I_{(\varepsilon_i, \varepsilon_{i+1}, \ldots, \varepsilon_{i+m-1})=B} \quad (1 \leq j \leq N),$$

$$T_{templ} = \sum_{j=1}^{N} \frac{(W_j - \mathbf{E}W_j)^2}{\mathbf{D}W_j}$$

Limit distributions of separate statistics $T_{mon}, T_{fr}, T_{templ}, T_{runs}$ are given in NIST STS.

The following results were obtained by V.G. Mikhailov in 2019 – 2020.

The limit distribution of the vector $\left(T_{mon}, T_{runs}, \frac{\pi_1 - \frac{M}{2}}{\sqrt{n}}, \ldots, \frac{\pi_N - \frac{M}{2}}{\sqrt{n}}\right)$ was obtained and asymptotic behavior (as $n \to \infty$) of its moments was described. It was shown that:

1) (T_{mon}, T_{fr}) and T_{runs} are asymptotically independent,

2) T_{mon} and T_{fr} are asymptotically uncorrelated.

Some close results

V.G. Mikhailov (2019–2020): $\left(T_{mon}, T_{fr}, T_{runs}\right)$.

A.A. Serov (2020): formulas for the numbers of sequences containing a given pattern a given number of times. By means of these results one may find the distribution of statistics of the NIST overlapping matching test for binary sequences and arbitrary pattern parameters.

Theorem

Let $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ be a sequence of independent Bernoulli trials with probability of success $\frac{1}{2}$. Suppose that random vectors $\vec{\eta}^{(i)} = (\eta_1^{(i)}, \eta_2^{(i)}, \eta_3^{(i)})$, $1 \le i \le N$, are indepedent and $\vec{\eta}^{(i)} \sim \mathcal{N}(0, C)$. If $m \ge 2$, N and B are fixed, then

$$\mathcal{L}\left(\left(\sqrt{N}T_{mon}, T_{fr}, \frac{2^{-m} - (2m-1)4^{-m}}{a^2} T_{templ}, \sqrt{N}T_{runs}\right)\right) \xrightarrow{d}$$

$$\xrightarrow{d} \sum_{i=1}^{N} (\eta_1^{(i)}, (\eta_1^{(i)})^2, (\eta_2^{(i)})^2, \eta_3^{(i)}), \quad n \to \infty.$$

The covariance matrix of the vector $\sum_{i=1}^{N} (\eta_1^{(i)}, (\eta_1^{(i)})^2, (\eta_2^{(i)})^2, \eta_3^{(i)})$ is

$$F = N \cdot \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 2C_{1,2}^2 & 0 \\ 0 & 2C_{1,2}^2 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Lemma

Let $\varepsilon_1, \varepsilon_2, \ldots$ be a sequence of independent Bernoulli trials with probability of success $p \in (0,1)$. If $p \neq \frac{1}{2}$, then
$(T_{mon}, T_{fr}) \xrightarrow{a.s.} (\mathrm{sgn}(2p-1) \cdot (+\infty), +\infty)$. *If $p_B \neq 2^{-m}$, then $p \neq \frac{1}{2}$ and*

$$(T_{mon}, T_{fr}, T_{templ}) \xrightarrow{a.s.} (\mathrm{sgn}(2p-1) \cdot (+\infty), +\infty, +\infty).$$

For all $p \in (0,1)$ we have $T_{runs} \xrightarrow{d} \xi$, where $\xi \sim N(0,1)$.

Corollary

Let $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ be a sequence of independent Bernoulli trials with probability of success $\frac{1}{2}$. Let $m \geq 2$, N and B be fixed and $n \to \infty$.

- *The limiting distribution of $(T_{mon}, T_{fr}, T_{templ}, T_{runs})$ is such that any two of its components, except for the 2nd and 3rd ones, are pairwise uncorrelated. The 2nd and the 3rd components are uncorrelated if and only if $\sum_{i=1}^{m} b_i = \frac{m}{2}$.*
- *If $\sum_{i=1}^{m} b_i = \frac{m}{2}$, then (T_{mon}, T_{fr}) and T_{templ} are asymptotically independent.*
- *If at least one of the statistics T_{mon}, T_{fr} is asymptotically independent of T_{templ}, then $\sum_{i=1}^{m} b_i = \frac{m}{2}$.*
- *The number of sign changes in pattern B is equal to half of the maximum possible, i.e. $\sum_{i=1}^{m-1} I_{b_i \neq b_{i+1}} = \frac{m-1}{2}$, if and only if the statistics T_{templ} and T_{runs} are asymptotically independent.*
- *T_{mon} and T_{fr} are asymptotically uncorrelated and asymptotically dependent.*
- *(T_{mon}, T_{fr}) and T_{runs} are asymptotically independent.*

Recommendation:

A necessary and sufficient condition for independence of the vector (T_{mon}, T_{fr}) and the statistics T_{templ} is $\sum_{i=1}^{m} b_i = \frac{m}{2}$. Hence it makes sense to use a template B containing exactly half of ones.

«Serial Test»

Denote by $\nu_{i_1 \ldots i_m}$ the frequency of the template $(i_1 \ldots i_m) \in \{0,1\}^m$ in the «cyclic» sequence $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{n-1}, \varepsilon_n, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{m-1}$. In other words,

$$\nu_{i_1 \ldots i_m} = \sum_{i=1}^{n} I_{(\tilde{\varepsilon}_i \ldots \tilde{\varepsilon}_{i+m-1})=(i_1 \ldots i_m)},$$

where $\tilde{\varepsilon}_i = \varepsilon_i I_{i \le n} + \varepsilon_{i-n} I_{i>n}$. Put

$$\psi_m^2 = \frac{2^m}{n} \sum_{i_1 \ldots i_m} \left(\nu_{i_1 \ldots i_m} - \frac{n}{2^m} \right)^2 = \frac{2^m}{n} \sum_{i_1 \ldots i_m} \nu_{i_1 \ldots i_m}^2 - n$$

Two statistics of «Serial Test»:

$$T_{serial1}(m) = \psi_m^2 - \psi_{m-1}^2, \; T_{serial2}(m) = T_{serial1}(m) - T_{serial1}(m-1).$$

Note that $T_{serial1}(m) = \sum_{j=1}^{m} T_{serial2}(j)$.

«Serial Test» and «Approximate Entropy Test»

Put

$$\Phi_m = \sum_{i_1 \dots i_m} \frac{\nu_{i_1 \dots i_m}}{n} \ln\left(\frac{\nu_{i_1 \dots i_m}}{n}\right), \quad T_{entr} = 2n\Big(\Phi_{m+1} - \Phi_m + \ln 2\Big).$$

Statistics $T_{serial1}(m)$ and $T_{serial2}(m)$ are used in «Serial Test», and the statistic T_{entr} is used in «Approximate Entropy Test».

Papers:
Kiyevets N.G., Korzun A.I (2014)
Voloshko V. A., Kharin Yu. S., Trubey A. I. (2022),
Voloshko V. A., Trubey A. I. (2022),
Voloshko V. A. (2023),
Savelov M. P. (2023).

«Serial Test» and «Approximate Entropy Test»

1) $T_{entr}(m)$ and $T_{serial1}(m+1)$ are asymptotically equivalent;
2) $T_{serial2}(m_1), T_{serial2}(m_2), \ldots, T_{serial2}(m_k)$ are asymptotically independent for any $m_1 < m_2 < \ldots < m_k \Rightarrow T_{serial1}(m)$ is a sum of asymptotically independent statistics:

$$T_{serial1}(m) = \sum_{j=1}^{m} T_{serial2}(j);$$

3) $T_{entr}(m_1)$ and $T_{serial2}(m_3)$ are asymptotically independent
$\Leftrightarrow m_3 > m_1 + 1$;
4) $T_{entr}(m_1)$ and $T_{serial1}(m_2)$ are asymptotically dependent;
5) $T_{serial1}(m)$ and $T_{serial2}(m)$ are asymptotically dependent.

Hence NIST package uses a triplet of asymptotically dependent statistics $T_{entr}(m_1), T_{serial1}(m), T_{serial2}(m)$.

«Serial Test» and «Approximate Entropy Test»

Recommendation:

In «Serial Test» and «Approximate Entropy Test» the statistics $T_{entr}(m_1)$, $T_{serial1}(m)$, $T_{serial2}(m)$ are used. Instead of these three statistics we recommend to use a triplet of statistics $T_{entr}(m_1)$, $T_{serial2}(m_2)$, $T_{serial2}(m_3)$, whose parameters satisfy the inequality $m_3 > m_2 > m_1 + 1$, because these three statistics are asymptotically independent. For similar reasons, it can be recommended to use a triple of statistics $T_{serial1}(m_1)$, $T_{serial2}(m_2)$, $T_{serial2}(m_3)$, the parameters of which satisfy the inequality $m_3 > m_2 > m_1$.

«Test for the Longest Run of Ones in a Block»

Fix L. We split $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ into non-overlapping blocks of length L. The first block has the form $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_L$, the second one has the form $\varepsilon_{L+1}, \varepsilon_{L+2}, \ldots, \varepsilon_{2L}$, and so on. The number of blocks

$$Q = \left[\frac{n}{L}\right].$$

We suppose: L is fixed, $n \to \infty$.
j-th block $(1 \le j \le Q)$ has the form $(\varepsilon_{(j-1)L+1}, \varepsilon_{(j-1)L+2}, \ldots, \varepsilon_{jL})$. Let ζ_j the length of the longest sequence of consecutive ones in this block:

$$\zeta_j = \max\left(s \in \{0, 1, \ldots, L\} \Big| \prod_{r=1}^{s} \varepsilon_{(j-1)L+i+r} = 1 \text{ for some } \ 0 \le i \le L - s\right).$$

Random variables ζ_j, $1 \le j \le Q$, Q, are independent and identically distributed; $0 \le \zeta_j \le L$.

Fix $K \geq 1$ and partition the set $\{0, 1, \ldots, L\}$ into $K+1$ non-empty disjoint subsets: $\{0, 1, \ldots, L\} = \bigsqcup_{j=0}^{K} \alpha_j$. Put

$$\nu_k = \sum_{i=1}^{Q} I_{\zeta_i \in \alpha_k}, \quad 0 \leq k \leq K,$$

$$w_k = \mathbf{P}(\zeta_j \in \alpha_k), \quad 0 \leq k \leq K.$$

Consider

$$T_{longrun} = \sum_{k=0}^{K} \frac{(\nu_k - Qw_k)^2}{Qw_k}.$$

$$T_{mon} = \frac{2S_n - n}{\sqrt{n}}, \quad T_{fr} = 4M \sum_{j=1}^{N} \left(\pi_j - \frac{1}{2} \right)^2, \quad T_{longrun} = \sum_{k=0}^{K} \frac{(\nu_k - Qw_k)^2}{Qw_k(\frac{1}{2})}.$$

We suppose that numbers N, L, K and sets $\alpha_0, \ldots, \alpha_K$ are fixed and $n \to \infty$. It's known that

$$\mathcal{L}(T_{mon}) \xrightarrow{d} \mathcal{N}(0,1), \quad \mathcal{L}(T_{fr}) \xrightarrow{d} \chi^2_N, \quad \mathcal{L}(T_{longrun}) \xrightarrow{d} \chi^2_K$$

as $n \to \infty$.

Theorem (2022)

Let $\varepsilon_1, \varepsilon_2, \ldots$ be a Bernoulli sequence with parameter $\frac{1}{2}$. Let $\vec{Y}^{(i)} = (Y_1^{(i)}, Y_2^{(i)}, \ldots, Y_{K+2}^{(i)})$, $1 \le i \le N$, be independent random vectors having the distribution $\mathcal{N}(0, C)$. If numbers N, L, K and sets $\alpha_0, \ldots, \alpha_K$ are fixed, then

$$\left(\sqrt{N}T_{mon}, T_{fr}, NT_{longrun}\right) \xrightarrow{d} \left(\sum_{i=1}^{N} Y_{K+2}^{(i)}, \sum_{i=1}^{N} \left(Y_{K+2}^{(i)}\right)^2, \sum_{j=1}^{K+1} \left(\sum_{i=1}^{N} Y_j^{(i)}\right)^2\right)$$

as $n \to \infty$. The covariance matrix of the limit vector $\left(T_{mon}, T_{fr}, T_{longrun}\right)$ has the form

$$F = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2N & 2\sum_{j=1}^{K+1} C_{j,K+2}^2 \\ 0 & 2\sum_{j=1}^{K+1} C_{j,K+2}^2 & 2K \end{pmatrix}.$$

Put $S = \sum_{j=1}^{K+1} C_{j,K+2}^2$.

Corollary

Let $\varepsilon_1, \varepsilon_2, \ldots$ be a Bernoulli sequence with parameter $\frac{1}{2}$, numbers N, L, K and sets $\alpha_0, \ldots, \alpha_K$ are fixed and $n \to \infty$.

1. *T_{mon} and $(T_{fr}, T_{longrun})$ are asymptotically uncorrelated, T_{fr} and $T_{longrun}$ are asymptotically nonnegatively correlated.*
2. *T_{fr} and $T_{longrun}$ are asymptotically uncorrelated $\Longleftrightarrow S = 0$.*
3. *T_{fr} and $T_{longrun}$ are asymptotically positively correlated $\Longleftrightarrow S \neq 0$.*
4. *T_{mon} and T_{fr} are asymptotically dependent.*
5. *If at least one of the statistics T_{mon} and T_{fr} is asymptotically independent of $T_{longrun}$, then $S = 0$.*
6. *If $S = 0$, then the vector (T_{mon}, T_{fr}) and the statistic $T_{longrun}$ are asymptotically independent.*

Example

Consider $L = 8$. NIST STS recommends to put

$$K = 3, \alpha_0 = \{0, 1\}, \alpha_1 = \{2\}, \alpha_2 = \{3\}, \alpha_3 = \{4, 5, 6, 7, 8\}.$$

In this case $S > 0$ and $T_{mon}, T_{fr}, T_{longrun}$ are pairwise asymptotically dependent.

In order for the vector (T_{mon}, T_{fr}) to «become» asymptotically independent with the statistic $T_{longrun}$, it suffices to replace the number K and the sets $\alpha_0, \ldots, \alpha_K$ with those

$$L = 8, K = 1, \alpha_0 = \{2, 5, 8\}, \alpha_1 = \{0, 1, \ldots, L\} \backslash \alpha_0 \Rightarrow S = 0.$$

Recommendation:

For fixed numbers L and K it makes sense to use sets $\alpha_0, \ldots, \alpha_K$ such that $\sum_{j=1}^{K+1} C^2_{j,K+2} = 0$ (if such sets exist).

Alternative H_1

Consider the following alternative H_1.
Let us split the sequence $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ into N disjoint blocks of length $L = \left[\frac{n}{N}\right]$. First block: $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_L)$, second block: $(\varepsilon_{L+1}, \varepsilon_{L+2}, \ldots, \varepsilon_{2L})$, etc. Let us number templates of length s of zeros and ones in lexicographic order: $B_1^{(s)} = (00\ldots0), \ldots, B_{2^s}^{(s)} = (11\ldots1)$. Let $\nu_k^{(N,s,h)}\left(B_j^{(s)}\right)$ be the normalized number of occurrences of the template $B_j^{(s)}$ among all chains of the form $(\varepsilon_{hi+1}, \varepsilon_{hi+2}, \ldots, \varepsilon_{hi+s})$ located in the k-th block:

$$\nu_k^{(N,s,h)}\left(B_j^{(s)}\right) = \frac{h}{L-s+h} \sum_{\frac{L(k-1)}{h} \le i \le \frac{Lk-s}{h}} I_{(\varepsilon_{hi+1}, \varepsilon_{hi+2}, \ldots, \varepsilon_{hi+s}) = B_j^{(s)}}$$

for $1 \le j \le R^s$ and $1 \le k \le N$. Consider vectors

$$\nu_k^{(N,s,h)} = \left(\nu_k^{(N,s,h)}\left(B_1^{(s)}\right), \nu_k^{(N,s,h)}\left(B_2^{(s)}\right), \ldots, \nu_k^{(N,s,h)}\left(B_{2^s}^{(s)}\right)\right),$$

$$a(s) = \left(2^{-s}, 2^{-s}, \ldots, 2^{-s}\right).$$

If H_0 is true, then the statistics $\sqrt{L}\left(\nu_1^{(N,s,h)} - a(s)\right), \sqrt{L}\left(\nu_2^{(N,s,h)} - a(s)\right), \ldots, \sqrt{L}\left(\nu_N^{(N,s,h)} - a(s)\right)$ are asymptotically independent and

$$\mathcal{L}\left(\sqrt{\frac{L}{h}}\left(\nu_k^{(N,s,h)} - a(s)\right)\right) \to \mathcal{N}(0, C(s,h)), \quad n \to \infty, \tag{1}$$

An alternative H_1: a scheme of series, such that in the nth series n random variables $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are considered and the following conditions are fulfilled.

Condition I(N, s, h)**.** The statistics

$$\sqrt{L}\left(\nu_1^{(N,s,h)} - a(s)\right), \sqrt{L}\left(\nu_2^{(N,s,h)} - a(s)\right), \ldots, \sqrt{L}\left(\nu_N^{(N,s,h)} - a(s)\right)$$

are asymptotically independent.

Condition II$\left(N, s, h, \boldsymbol{\mu}^{(N,s)}\right)$. For all $1 \le k \le N$

$$\mathcal{L}\left(\sqrt{\frac{L}{h}}\left(\nu_k^{(N,s,h)} - a(s)\right)\right) \to \mathcal{N}\left(\mu_k^{(N,s)}, C(s,h)\right), \quad n \to \infty,$$

where the matrix $C(s,h)$ is the same as in (1).

Examples of H_1 satisfying the conditions $\mathbf{I}(s)$ and $\mathbf{II}\left(s, \boldsymbol{\mu}^{(s)}\right)$:

- H_0
- $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent random variables having the Bernoulli distribution $Bern\left(\frac{1}{2} + \frac{\delta_n}{\sqrt{n}}\right)$, where δ_n is some numerical sequence that has a finite limit.
- For each n, the sequence $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ is a homogeneous strictly stationary Markov chain, which, as n grows, «converges» to the sequence of independent random variables with distribution $Bern(\frac{1}{2})$ Namely:

Let $r \geq 1$. Alternative $\tilde{H}_1$:
firstly, for each n the sequence $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ is the first n elements of a homogeneous strictly stationary Markov chain $\{X_n(t), t \geq 1\}$ of order $r-1$, and secondly,

$$\mathbf{P}_n^{\tilde{H}}(\varepsilon_r = i_r | \varepsilon_1 = i_1, \ldots, \varepsilon_{r-1} = i_{r-1}) = \frac{1}{2}\left(1 + \frac{\delta^{(r)}(i_1, \ldots, i_r) + o(1)}{\sqrt{L}}\right)$$

as $n \to \infty$, where the quantities r and $\delta^{(r)}(i_1, \ldots, i_r)$ do not depend on n. Voloshko V. A., Kharin Yu. S., Trubey A. I. (2022); Voloshko V.A., Trubey A.I. (2022), Voloshko V. A. (2023).

Another example, when the conditions $\mathbf{I}(s)$ and $\mathbf{II}\left(s, \boldsymbol{\mu}^{(s)}\right)$ are fulfilled: the tested sequence is obtained as "gluing" of several homogeneous strictly stationary Markov chains

Savelov M.P. (2024)

Corollary

If the conditions $\mathbf{I}(N, s, h)$ *and* $\mathbf{II}\left(N, s, h, \boldsymbol{\mu}^{(N,s)}\right)$ *are fulfilled, the parameters of statistics* T_{mon}, T_{fr}, T_{runs}, $T_{longrun}$, T_{matrix}, T_{templ}, $T_{lincompl}$, $T_{serial1}$, $T_{serial2}$, T_{entr} *are fixed and some conditions on* N, s, h *are fulfilled, then*

$$\left(T_{mon}, T_{fr}, T_{runs}, T_{longrun}, T_{matrix}, T_{templ}, T_{lincompl}, T_{serial1}, T_{serial2}, T_{entr}\right)$$

converges in distribution to a vector whose components are linear or quadratic forms of the components of some gaussian vector with a known distribution.

Similar results were obtained for generalizations of statistics of NIST tests in the case when H_0 corresponds to a polynomial scheme and H_1 is «close» to H_0.

Recommendations:

1. It's better to use $T_{Fourier}(\ln 20, \gamma^*)$ or $T^{\varepsilon}_{Fourier}$ for some $\varepsilon \in (0, \frac{1}{10}]$ instead of $T_{Fourier}(\ln 20, 4)$ in Discrete Fourier Transform (Spectral) Test.
2. A necessary and sufficient condition for independence of the vector (T_{mon}, T_{fr}) and the statistics of «Non-overlapping Template Matching Test» is $\sum_{i=1}^{m} b_i = \frac{m}{2}$. Hence it makes sense to use a template B containing exactly half of ones.
3. Instead of $(T_{entr}(m_1), T_{serial1}(m), T_{serial2}(m))$ we recommend to use $(T_{entr}(m_1), T_{serial2}(m_2), T_{serial2}(m_3))$ such that $m_3 > m_2 > m_1 + 1$, or $(T_{serial1}(m_1), T_{serial2}(m_2), T_{serial2}(m_3))$ such that $m_3 > m_2 > m_1$.
4. A necessary and sufficient condition for independence of the vector (T_{mon}, T_{fr}) and the statistic $T_{longrun}$ is $\sum_{j=1}^{K+1} C^2_{j,K+2} = 0$. Hence for fixed numbers L and K it makes sense to use sets $\alpha_0, \ldots, \alpha_K$ such that $\sum_{j=1}^{K+1} C^2_{j,K+2} = 0$ (if such sets exist).

Full version

All results:
Savelov M.P., Diskr. Mat., 2021–2024

Thank you for attention!