Reliability of two-level testing approach of the NIST SP800-22 test suite and two-sided estimates for quantiles of binomial distribution

Aleksandr Serov

Academy of Cryptography of the Russian Federation
Steklov Mathematical Institute of Russian Academy of Sciences

Volgograd
June 9, 2023

Random sequences are used in a large variety of areas, such as *statistics*, *cryptography*, *science* in general, *game theory*, *gambling*, *quantum mechanics* and so on.

Random sequences may be generated either by *physical sources* or *deterministic algorithms*.

Random Number Generators (RNGs) represent a fundamental component in many applications, they are crucially important for cryptographic systems.

The security of main part of cryptographic schemes and protocols is based on the perfect randomness of RNG outputs, RNGs are classified into two types: PRNGs and TRNGs

The quality of RNG outputs is assessed by statistical tests, which evaluate whether the output sequences conform with the given *null hypothesis* $\mathcal{H}_0$ (e. g., the elements of the sequence are independent and uniformly distributed) or not.

Statistical tests are also used to evaluate the outputs of *block ciphers* and *hash functions*, to preliminarily validate the indistinguishability of their outputs from a *uniform random permutation* or a *equiprobable random mapping*.

The quality checking of binary sequences usually is based on some well-known batteries of tests, each of which is composed of several tests, e.g. DIEHARD, SP800-22, TESTU01.

Note that a test for randomness may be interpreted only in a probabilistic way.

Is the sequence $\underbrace{000000000000000000000000000000000000000}_{n=39}$ random?

From a mathematical point of view a test may be considered as a function of a sequence of n elements with output value in $[0, 1]$, called a p-value

$$p = \mathbf{P}\left(S \geq s \mid \mathcal{H}_0\right),$$

where S — statistic, s — observed statistic value, $\mathcal{H}_0$ — null hypothesis.

If $\mathcal{H}_0$ is true and the distribution of test statistic is continuous, then p-value has a uniform distribution over the interval $[0, 1]$.

If $\mathcal{H}_1$ (alternative hypothesis) is true then the distribution of the p-value is not known as a rule and determined by S distribution.

In the discrete case, then the test statistic distribution is not continuous

- if $\mathcal{H}_0$ is true, i. e. all elements of the input sequence are independent and uniformly distributed, then the p-value is *approximately uniformly distributed* in the interval $[0, 1]$ and its cumulative distribution function (CDF) $F_p(x)$ is *stepwise* and equal to x in discrete set of points.
- If $\mathcal{H}_1$ is true, i. e. the elements of the generated sequence is non-independent or not distributed according to the uniform distribution, then the CDF of p-value is not known as a rule.

The SP 800-22 test suite from the NIST, the most commonly used statistical test suite, is considered.

The main reason for using the suite is that it has several appealing properties:

- all tests of the suite are applied to each tested sequence of n bits, the result of each test is one or more p-values,
- the suite is composed by a number of well known tests,
- the source code of all the tests in the suite is public available.

The full version of NIST SP 800-22 test suite consists of 15 tests

#	Test Name	Used Statistics
1	*Frequency*	normalized modulus of the difference between frequencies of 1 and 0
2	*Block Frequency*	χ^2 statistics of 1 frequencies in adjacent non-intersecting 128-bit blocks
3	*Cumulative Sums*	maximal deviation from 0 for partial sums of ± 1 walk constructed by the binary sequence
4	*Runs*	the total number of 1-runs and 0-runs
5	*Longest Run*	χ^2 statistics of maximal lengths of 1-runs frequencies in adjacent non-intersecting 10^4-bit blocks
6	*Binary Matrix Rank*	χ^2 statistics of binary 32×32-matrices ranks of frequencies formed from adjacent nonintersecting 1024-bit blocks of the sequence
7	*Discrete Fourier Transform*	normalized difference between the number of DFT coefficients of n-bit sequence, exceeding $\sqrt{n \log 20}$ in absolute value, and $0.95n/2$

#	Test Name	Used Statistics
8	*Overlapping Templ. Match.*	χ^2 statistics of 9-bit 1-series frequencies in 1032-bit adjacent blocks with overlappings
9	*Universal statistical test*	sum of base 2 logarithms of distances between equal nonintersecting 7-bit blocks
10	*Approximate Entropy*	difference of logarithmic frequencies statistics of 10- and 11-bit segments with overlappings, the reference distribution is the χ^2 distribution
11	*Serial*	differences χ^2 statistics of frequencies of all 14-, 15- and 16-bit segments with overlappings
12	*Linear Complexity*	χ^2 statistics of shortest LSR lengths frequencies generating 500-bit segments of the sequence
13	*Non-overlap. Templ. Match.*	χ^2 statistics of aperiodic fixed segments frequencies in 8 adjacent non-overlapping blocks
14	*Random Excursion*	χ^2 statistics of Random Walk cycles frequencies with fixed numbers visiting of $-4, -3, \ldots, 4$ states
15	*Random Excursion Variant*	the frequencies of visiting 18 states from -9 to 9 by Random Walk

Let us describe the testing strategy suggested in NIST SP 800-22 publication and discuss under which assumptions this strategy increases the reliability and when produces incorrect results.

The whole bit sequence is divided into N blocks of n bits each.

For each value of test statistic the p-value is computed according to the theoretic test statistic distribution

If p-value is larger than the level of significance α, then we say that the tested block passes the test.

For each test the set of N computed p-values is further tested by the *first-level* and the *second-level* tests.

The first-level test is based on the fraction of blocks passed the test.

The second-level test checks the uniformity of the p-values distribution.

If RNG produces sequences for which the hypothesis $\mathcal{H}_0$ holds, then we will call it *ideal RNG*

If $\mathcal{H}_0$ is true and $p \leq \alpha$, where α — *level of significance*, then we have *Type I error* and may compute its probability

$$\mathbf{P}_{\mathcal{H}_0}\{p \leq \alpha\} = F_p^{(0)}(\alpha) = \alpha.$$

If $\mathcal{H}_0$ is false and $p > \alpha$, then we accept the sequence as random and have *Type II error*, denote the probability of this event by β.

So we can commit two errors:

I) reject $\mathcal{H}_0$, when the sequence is generated by an ideal RNG,

II) accept $\mathcal{H}_0$, when the sequence is generated by a non-ideal RNG.

The approach when we compare p-value with α only is known as *first-level testing* approach.

The second-level testing approach has been known for a long time and it may increase the testing power as compared to the first-level approach:

(a) for the computed sequence of N p-values for a particular statistical test, compute the fraction ζ of the sequences passed the test. For example,

$$\zeta = m/N,$$

where m is the number of sequences with p-values $\geq$ than α.

If $\mathcal{H}_0$ is true, then

$$\zeta N \sim \mathrm{Bin}(N, 1-\alpha).$$

If N is large enough, the ζ distribution may be approximated by $\mathcal{N}(\mu, \sigma^2)$, where

$$\mu = 1-\alpha \quad \text{and} \quad \sigma = \sqrt{\alpha(1-\alpha)/N}.$$

The range of acceptable values of fractions may be determined by the interval

$$\left[1-\alpha-3\sqrt{\frac{\alpha(1-\alpha)}{N}}, 1-\alpha+3\sqrt{\frac{\alpha(1-\alpha)}{N}}\right].$$

(b) to test the hypothesis on the uniformity of p-values distribution the interval $[0, 1]$ is divided into k disjoint sub-intervals, the number π_i of p-values falling in i-th sub-interval are counted and the statistic of chi-square goodness-of-fit test is computed:

$$\chi^2_{obs} = \sum_{i=1}^{k} \frac{(\pi_i - N/k)^2}{N/k}.$$

Uniformity checking is accomplished via an application of a *chi-square goodness-of-fit test* in k sub-interval with statistic χ^2. This statistical test yields a level-two p-value $p_{II} = \mathbf{P}_{\mathcal{H}_0}\left\{\chi^2 > \chi^2_{obs}\right\}$, which is calculated as follows

$$p_{II} = \frac{1}{\Gamma\left(\frac{k-1}{2}\right)} \int_{\chi^2_{obs}/2}^{\infty} t^{(k-1)/2-1} e^{-t} dt, \quad \Gamma(z) = \int_0^{\infty} t^{z-1} e^{-t} dt.$$

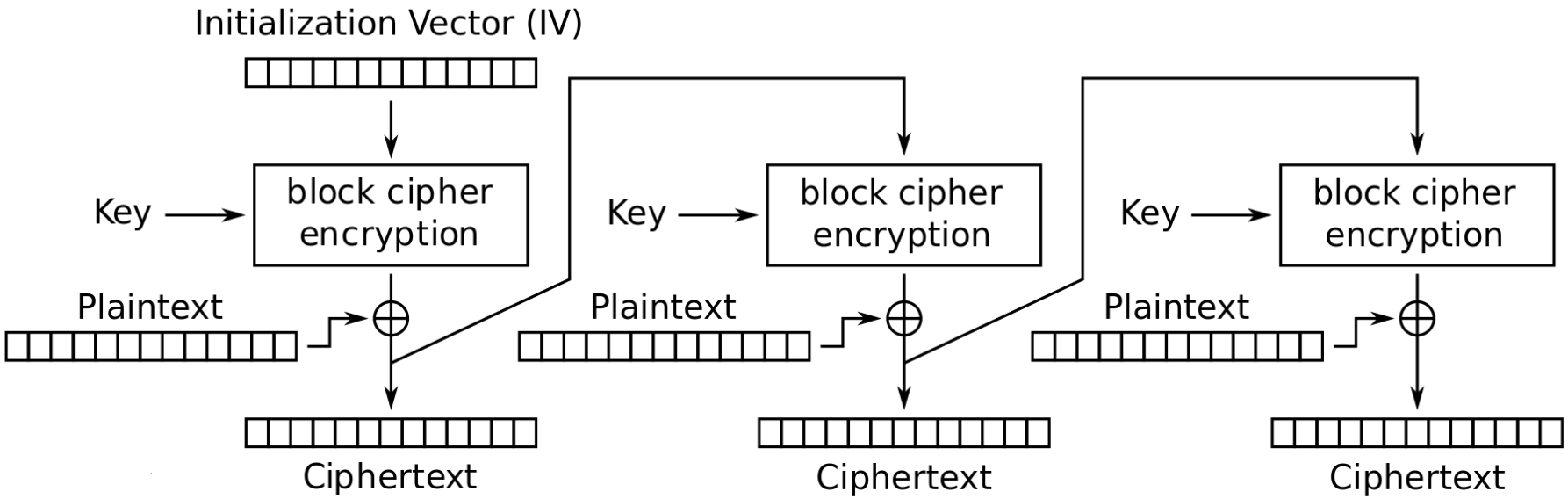
Given a significance α_{II}, $\mathcal{H}_0$ is rejected if $p_{II} \leq \alpha_{II}$, otherwise the sequences may be considered to be uniformly distributed. In SP 800-22 NIST recommends to use $\alpha_{II} = 0.0001$ and $k = 10$ sub-intervals.

The choice of N in a two-level approach is usually a tradeoff.

It was shown that for extremely large values of N the level-two approach always fail.

All 15 tests from NIST SP 800-22 test suite were applied to pseudorandom sequences generated by AES block cipher.

For generation such sequences AES block cipher in the CFB (Cipher Feedback Block) mode with the zero initialization vector (IV) and plaintext block (Plaintext) was used; the key for each sequence was randomly selected from the set of 128-bit binary sequences, the length in bits of all sequences was chosen to be the same and equal to $n = 2^{20} = 1.048.576$ each.



Using the AES-based PRNG allows to arbitrarily increase N.

Results of two-level test for the AES-based RNG

#	Test Name	χ^2 test, N sequences $N = 10^3$	$N = 10^4$	$N = 10^5$	$N = 2^{20}$
1	*Frequency*	0.616305	0.290806	0.588411	**0.000803**
2	*Block Frequency*	0.187581	0.773212	0.374097	**0.000125**
3	*Cumulative Sums*	0.401199	0.124765	0.959543	**0.000009**
4	*Runs*	0.150340	0.885418	0.910568	0.107966
5	*Longest Run*	0.610070	0.239883	**0.000355**	**0.000000**
6	*Binary Matrix Rank*	0.878618	0.341017	**0.000000**	**0.000000**
7	*Discrete Fourier Transform*	0.371941	0.014836	**0.000000**	**0.000000**
8	*Overlapping Templ. Match.*	0.071177	0.202268	**0.000000**	**0.000000**
9	*Universal statistical test*	0.574903	0.108534	**0.000000**	**0.000000**
10	*Approximate Entropy*	0.246750	0.078038	0.219501	**0.000000**
11	*Serial*	0.942198	0.174057	0.213964	0.572679
12	*Linear Complexity*	0.839507	0.279152	0.299852	0.117305
13	*Non-overlap. Templ. Match.*	0.092041	0.372782	0.121382	0.275416
14	*Random Excursion*	0.914727	0.663838	0.346173	**0.000028**
15	*Random Excursion Variant*	0.238264	0.133576	**0.000080**	**0.000000**

Let $\varepsilon = \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ be the tested sequence of bits of length n. Denote by

$$S_n = 2(\varepsilon_1 + \ldots + \varepsilon_n) - n = X_1 + \ldots + X_n,$$

where $X_i = 2\varepsilon_i - 1$, $i = 1, 2, \ldots, n$. If $\mathcal{H}_0$ is true, then $\frac{S_n+n}{2}$ to follow $\mathrm{Bin}\left(n, \frac{1}{2}\right)$ and by the classic De Moivre-Laplace theorem, for a sufficiently large n, the distribution of

$$\frac{\frac{S_n+n}{2} - \frac{n}{2}}{\sqrt{\frac{n}{4}}} = \frac{S_n}{\sqrt{n}} \quad \text{is closely approximated by } \mathcal{N}(0,1)$$

and

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

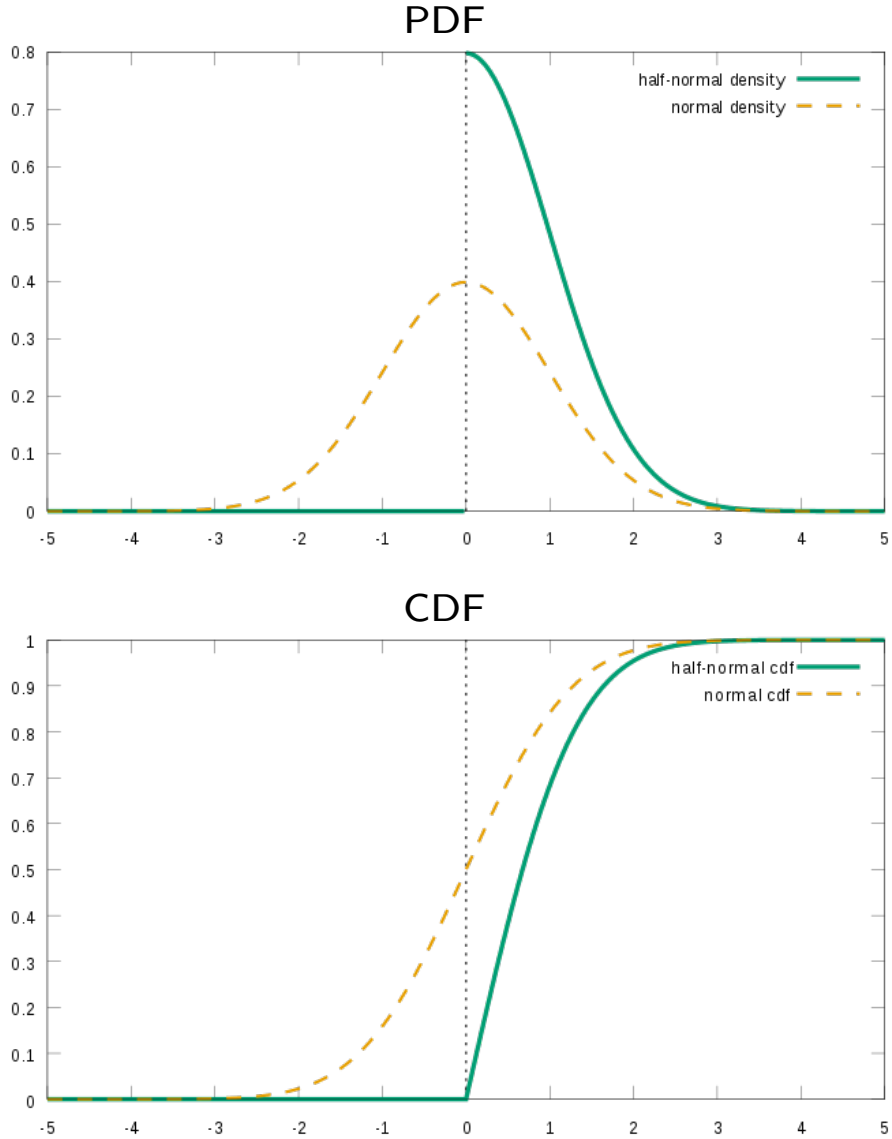
If X is normal, then $|X|$ is half normal and

$$F_{|X|}(z) = \mathbf{P}\left\{|X| \le z\right\}, \ z \ge 0, \quad F_{|X|}(z) = 2\Phi(z) - 1.$$

It implies that, the corresponding p-value of the test statistic $S(\varepsilon) = \frac{|X_1+\ldots+X_n|}{\sqrt{n}}$ is equal to

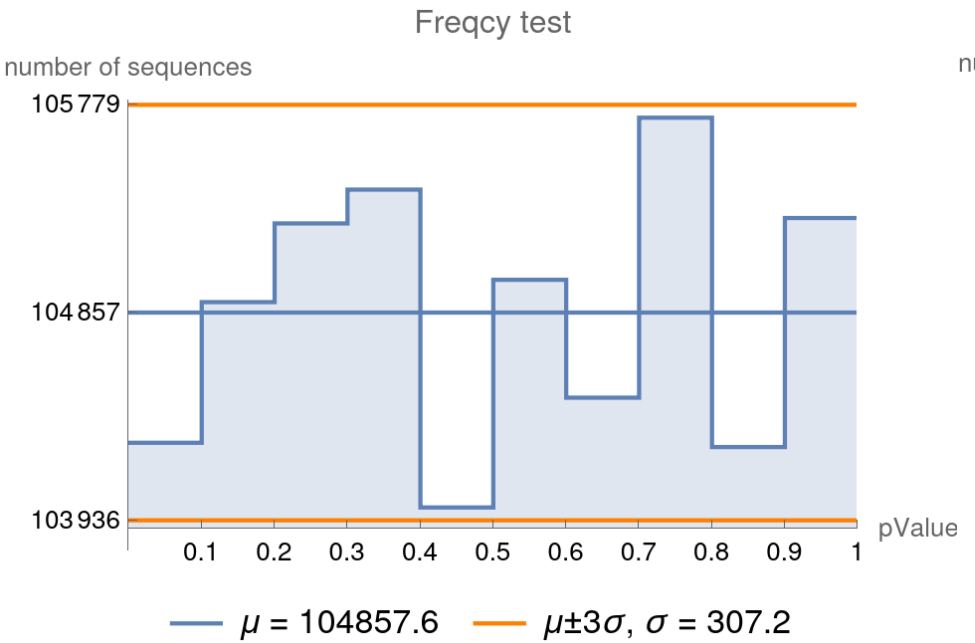
$$p_I = \lim_{n\to\infty} \left(1 - F_{|S_n|}(S(\varepsilon))\right) = 2\left(1 - \Phi(S(\varepsilon)\right).$$

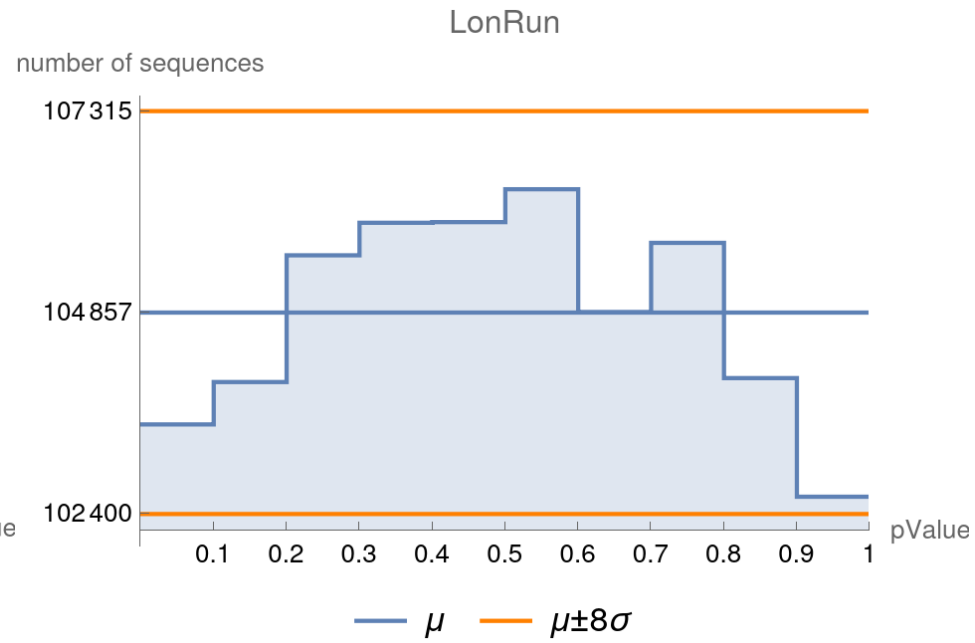
PDF and CDF of the half-normal distribution



$X \sim \mathcal{N}(0, \sigma^2)$, then $\mu_Y = \sqrt{\frac{2}{\pi}}\sigma$, $\sigma_Y = \sigma^2\left(1 - \frac{2}{\pi}\right)$ for $Y = |X|$.

Histograms of the sets of p-values

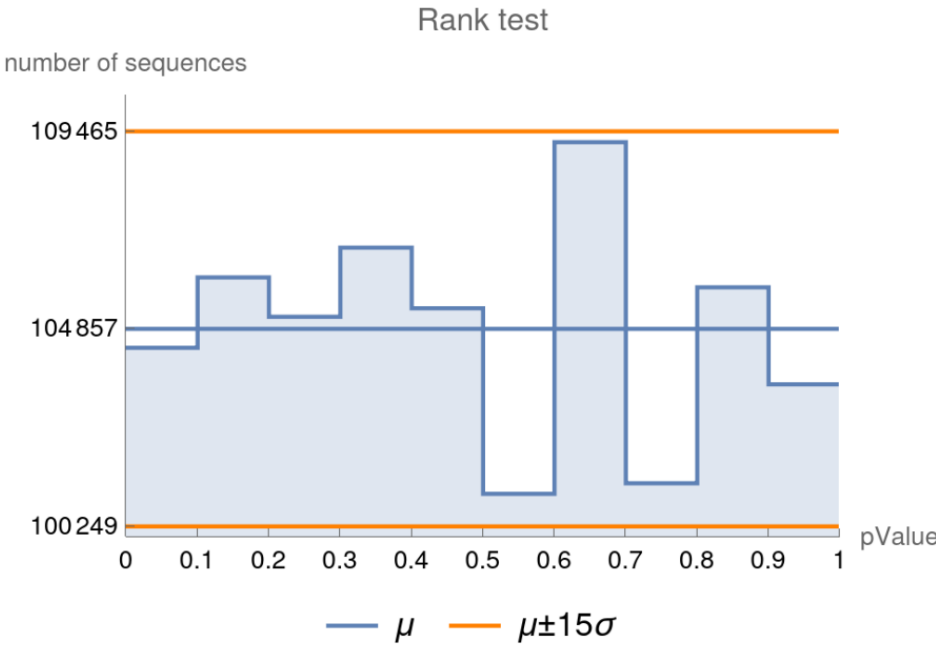


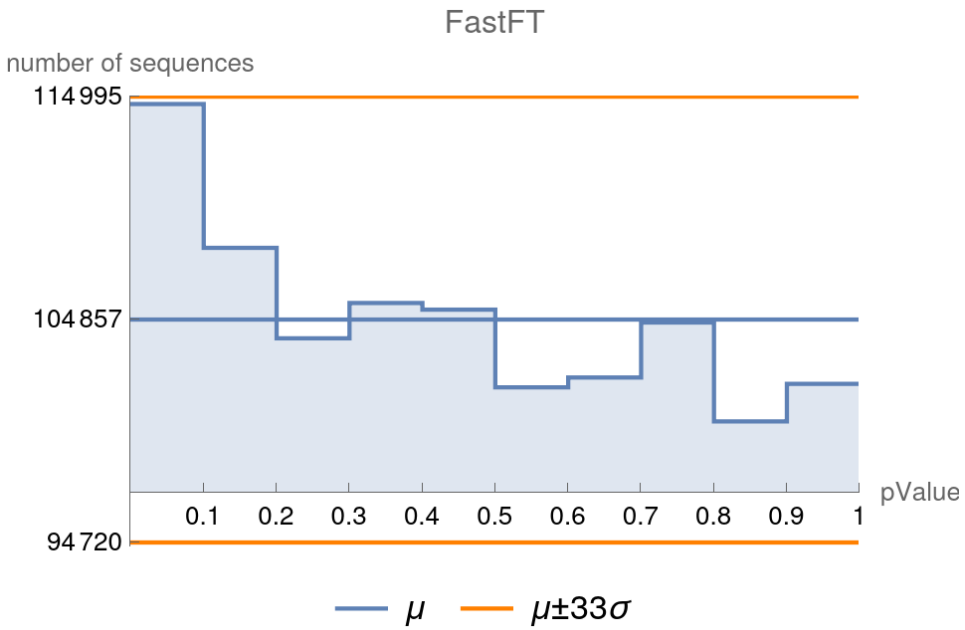


Comparison between expected deviation and measured deviation in the distribution of p-values in the interval $[0, 1]$ for Frequency test and Longest Run test

$$\sigma = \sqrt{N\left(1 - \frac{1}{k}\right)\frac{1}{k}}$$

Histograms of the sets of p-values



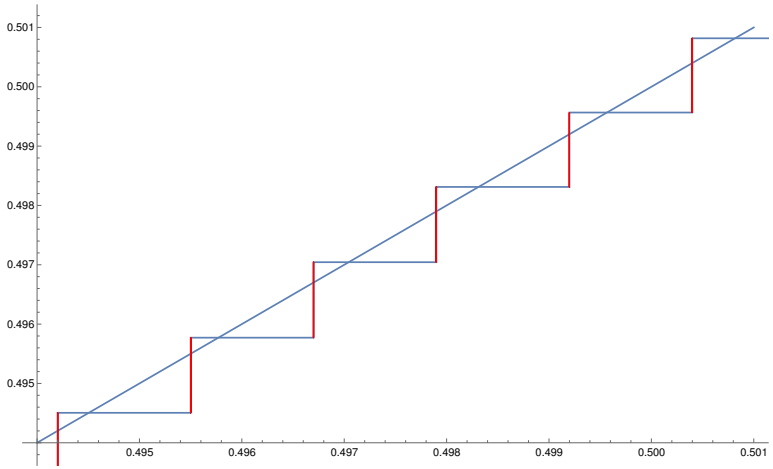


Comparison between expected deviation and measured deviation in the distribution of p-values in the interval $[0, 1]$ for Binary Matrix Rank test and Discrete Fourier Transform test

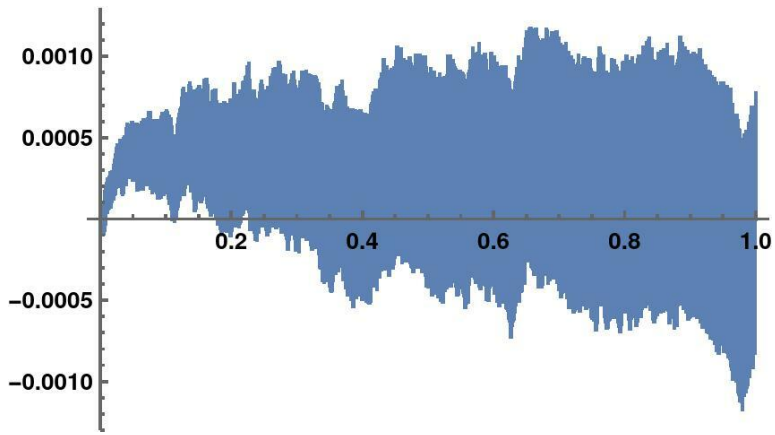
Let $p_1, \ldots, p_N \in [0,1]$ be p-values for AES-based PRNG, then the empirical CDF $F_N(x)$ for this sample is defined as follows

$$F_N(x) = \frac{1}{N} \sum_{j=1}^{N} H\left(x - p_j\right),$$

where $H(x)$ is the unity-step function, defined as $H(x) = \begin{cases} 0, & x < 0, \\ 1, & x \geq 0. \end{cases}$



(a)



(b)

(a) Comparison between empirical CDF F_N of the p-values for Frequency test applied to the AES-based sequences and the continuous uniform CDF, (b) differences between $F(x) = x$ and $F_N(x)$, $n = N = 2^{20}$.

Let $x_\alpha(n,p)$ be α-level quantile of the binomial distribution with parameters (n,p), i. e.

$$x_\alpha(n,p) = \min\{k\colon \mathbf{P}\{X_{n,p} \le k\} \ge \alpha\} \in \{0,1,\dots,n\},$$
$$x_0(n,p) = 0 \text{ and } x_1(n,p) = n,\ x_{1/2}(n,p) = \lfloor n/2 \rfloor.$$

Theorem. *The following estimates are true for* $0 < \alpha < 1/2$

$$r_1 - 1 \le x_\alpha(n,p) \le r_1,$$

where $r_1 = \left\lceil np + \frac{q-p}{6}\Phi^{-1}(1-\alpha)^2 - \Phi^{-1}(1-\alpha)\sqrt{npq + \frac{(q-p)^2}{6^2}\Phi^{-1}(1-\alpha)^2}\right\rceil$,

for $1/2 < \alpha \le 1$

$$r_2 - 1 \le x_\alpha(n,p) \le r_2,$$

where $r_2 = \left\lceil np + \frac{q-p}{6}\Phi^{-1}(\alpha)^2 + \Phi^{-1}(\alpha)\sqrt{npq + \frac{(q-p)^2\Phi^{-1}(\alpha)^2}{6^2}}\right\rceil$.

Corollary. *The following estimates are true*

$$r_1 - 1 \le x_\alpha(n,1/2) \le r_1, \quad 0 \le \alpha < 1/2, \quad \text{where } r_1 = \left\lceil \frac{n}{2} - \frac{\sqrt{n}}{2}\Phi^{-1}(\alpha)\right\rceil,$$

$$r_2 - 1 \le x_\alpha(n,1/2) \le r_2, \quad 1/2 < \alpha \le 1, \quad \text{where } r_2 = \left\lceil \frac{n}{2} + \frac{\sqrt{n}}{2}\Phi^{-1}(\alpha)\right\rceil.$$